July 2018

Investigation Repeater Effects on Small-sample Equating: Include or Exclude?

Hongyu Diao
University of Massachusetts-Amherst

INVESTIGATING REPEATER EFFECTS ON SMALL-SAMPLE EQUATING: INCLUDE OR EXCLUDE?

A Dissertation Presented

by

HONGYU DIAO

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2018

College of Education

Research, Educational Measurement, and Psychometrics

INVESTIGATING REPEATER EFFECTS ON SMALL-SAMPLE EQUATING: INCLUDE OR EXCLUDE?

A Dissertation Presented

By

HONGYU DIAO

Approved as style and content by:

_____
Lisa A. Keller, Chair

_____
Craig Wells, Member

_____
Anna Liu, Member

_____
Jennifer Randall
Associate Dean of Academic Affairs
College of Education

DEDICATION

To my fiancé Dr. Shengsheng Xu and my family

ACKNOWLEDGMENT

Firstly, I would like to express my sincere gratitude to my advisor, committee chair and my life mentor Professor Lisa Keller for her invaluable guidance, advice and continuous support during my Ph.D. career. She is my role model in career, academic and life.

I would like to thank the rest of my dissertation committee members: Professor Craig Wells and Professor Anna Liu for their insightful comments and questions which incented me to widen my research from various perspectives.

My sincere thanks also go to Professor Stephen Sireci. Without him, I was not able to start my academic journey and become a psychometrician. I would like to thank him for providing me the opportunity to study in REMP, being strict to my research and writing, and encouraging me for every little progress I made. I would also give a special thank to Professor Ronald Hambleton, who is the "grandpa" of my REMP family, for sharing his wisdom and knowledge to everybody all the time. In addition, I want to thank Professor Scott Monroe, Professor Jennifer Randall and Dr. April Zenisky for being patient to me and willing to answer all my silly questions.

I would thank all my brothers and sisters in my REMP family. To Francis, who is my cohort sister and my best buddy for being nice and sweet and positively influencing me in everything; To Fen, who offered me advice and snacks in the office; To HyunJoo, Yooyoung, Hwanggyu, Duy, Ella, Frank, Ale, Josh and Darius for bringing so much fun and laughs to the Furcolo.

Lastly, I would like to thank my parents for being understanding and giving me unconditional love; my friends Lili and Yunlu for our long-term friendship. In closing, I express my most profound gratitude to my fiancé Shengsheng for his endless love, support and encouragement throughout my gradate education and dissertation work. My Ph.D. is dedicated to him.

ABSTRACT

INVESTIGATE REPEATER EFFECTS ON SMALL-SAMPLE EQUATING: INCLUDE OR EXCLUDE?

MAY 2018

HONGYU DIAO, B.A., BEIJING WUZI UNIVERSITY

M.Ed., UNIVERSITY OF MASSACHUSETTS AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Lisa A. Keller

In licensure testing programs, some examinees might attempt the test multiple times till they are satisfied with their final score, those who take the same test repeatedly are referred to as repeaters. Previous studies suggested that repeaters should be removed from the total sample before implementing equating procedures for two reasons: 1) repeater group is distinguishable from the non-repeater group and the total group, 2) repeaters may memorize anchor items and cause an item drift in common items in the non-equivalent anchor test (NEAT) design. However, removing repeaters might not be the best solution if the testing program only has a small number of examinees (e.g., teaching licensure tests with 20-30 examinee per test form). Excluding repeaters may cause an even smaller sample size and results in high bias and errors (Kolen and Brennan, 2014). Additionally, the population invariance property might not hold because of the differences between total sample group and repeater group. Therefore, three solutions were purposed to deal with repeaters effects in the current study, they are: 1) excluding repeaters, 2) including repeaters but removing problematic anchor items, 3) applying Rasch equating to capitalize on the invariance property.

The main purpose was to investigate which solution(s) can mitigate the negative repeater effects. The secondary purpose was to compare identity equating, nominal weight equating, circle-arc equating and Rasch equating with small, medium to large sample size levels on a mixed-format test. The data generation was manipulated by repeater ability levels, repeater proportions, the drift in anchor test due to exposure and sample size levels. Both purposes were evaluated by equating bias, equating errors and population invariance measures. Furthermore, the practical implications were discussed based on the accuracy of pass/fail decision. Lastly, the recommendations regarding appropriate repeater effects solutions and small sample equating techniques were made based on given test conditions.

The most important finding reveals the performance of repeater effect solutions and small-sample equating techniques highly depend on the anchor test. If the anchor was not drifted, retaining all repeaters can provide higher equating accuracy and decision accuracy than excluding repeaters. However, if anchor test was problematic and drifted due to exposure. Using circle-arc equating and identity equating or removing repeaters can significantly prevent high equating bias.

Finally, the study recommends removing repeaters if the drift is unknown. At the small sample size levels (i.e., N =20 and N =50), identity equating had the most satisfactory performance. At higher sample size levels, circle-arc equating provided the most stable equating results while nominal weight mean equating can minimize the violation to invariance property of equating. Rasch equating, however, is not applicable to size levels smaller than 300.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

1.1 Background

In educational assessments, multiple forms of tests might be administered to different groups of examinees for security or practical considerations. To make the multiple forms comparable across administrations, these forms are assumed to have been constructed with similar content domains and statistic characteristics. Although test developers make all efforts to construct multiple forms of the test parallel in test content and statistics, it is common that there are variations in test specifications and statistics across forms. However, reporting the incomparable scores may cause some negative consequences. For example, if the college admission decision is made based on the score of an assessment that has multiple administrations. Examinees who take the test with lower difficulty level are likely to have higher testing score than those who are administered by the test with the higher difficulty level. The interpretations and the use of the score would vary depending on the forms of the test and therefore violating the fairness of the admission decisions that are made based on one reported score. As a result, it is important to adjust the scores of multiple forms on the same scale and make them comparable. The procedure of this adjustment is referred to as *equating*.

According to Kolen and Brennan, "Equating is a statistical process that is used to adjust scores on test forms so that scores can be used interchangeably" (2014, pp.2). The equating procedure is carried out to place the scores obtained from test X and Y on the same metric by establishing the correspondence between the scores on X and Y (Hambleton, Swaminathan, & Rogers, 1991). Usually, Y is referred to as the

reference/base/old form while X is referred to as the new Form where the score will be rescaled to the Form Y metric. The purpose of equating procedure is to estimate an equating function that relates the equivalent equated score of new Form X to base Form Y. Typically, equating function produces a raw-to-raw score conversion table that lists the equivalent score of base form to an integer score of the new form. The equating procedure might be different across varied equating techniques but the final goal is the same. After the equating procedure, examinees participate different administrations of the test would obtain a score on the same scale. As a result, one important purpose of the equating procedure is to prevent unfairness results from the variability of score interpretations across test forms or test administrations.

1.1.1 Properties of Test Equating

Some important equating properties were proposed in the literature (Angoff, 1971; Lord, 1980; Petersen, Kolen & Hoover, 1989; Harris and Crouse, 1993; Dorans & Holland, 2000). In practice, these properties play important roles to ensure the scores on alternate forms are interchangeable after estimating the equating function.

1. The symmetry property: Score on Form X can be equated to Form Y, score on Y will be equated to Form X using the same equating approach (Lord, 1980).
2. Same specification property: Alternate forms are built with the same content and statistical specifications.
3. Equity property: This property holds if examinees with a given true score would have identical observed score distributions on base Form Y and rescaled Form X (Lord, 1980).

4. Reliability property: Alternate forms of tests should have the same reliability.
5. Population invariance property: The equating function should remain same regardless of the groups of examinees used to conduct equating (e.g., males and females).

All of the properties are crucial prerequisites to equating analysis. The invariance property requires the equating function should remain the same across sexes, ethnicities or subgroups with respect to other demographic variables. If the property fails, the linkage function is referred as a *concordance* on a given subgroup rather than equating though the computation of the linking function for concordance is same as the equating function (Dorans & Holland, 2000; Dorans, 2004). The current research mainly focuses on the property of the group invariance and investigates one specific subgroup that consists of examinees who retake the same test multiple times.

In reality, there is no test completely population invariant, the question is to which degree the absence of the invariance is acceptable or negligible so that equating procedure can still perform across all groups (Kolen, 2004). Previous research had fully explored the population invariance criteria to quantify the degree of equating invariance such as root mean squared difference (RMSD) and root expected mean squared difference (REMSD) (e.g., Doran & Holland, 2000; von Davier, Holland, & Thayer, 2004; Dorans, 2004; Dorans, Liu & Hammond, 2008). The descriptions of these invariance measures are discussed in Chapter II.

1.1.2 Equating Designs

Three equating designs are commonly used to collect data for equating: the single group design, the random group design, and the non-equivalent group anchor test (NEAT) design. In single group design, the same examinees respond to items on both forms, any difference in scores of two forms can be attributed to the difference in difficulty. The limitation of this equating design is that the performance can be impacted by the order of administration and the fatigue effects due to long testing time. In random group design, test takers are randomly assigned to the forms of administration. In practice, a popular way to randomly assign forms to examinees is using the spiraling procedure. Examinees seated next to another receive alternate forms at the same time, examinees that receive the same form are considered assigned to the same group. The difference in performance is assumed to be the difference in test difficulty. This design is popular in the application if test forms can be administered in one administration. However, releasing both old and new forms within an administration may arise some concerns in test security than releasing different test forms across different testing administrations. In addition, performing random group design requires a large number of examinees to hold the assumption of randomness.

NEAT design is more commonly used in practice because it allows testing programs to administer different forms of the test to different groups of examinees at different dates (Kolen & Brennan, 2014). In the NEAT design, new and reference test forms share a set of common items V, which is a “mini version” of the total test. If the total test score includes scores on anchor test, the set of common items is referred to as an internal anchor; otherwise, it is referred to as an external anchor. NEAT design allows

more flexibility in test administration but the use of this design requires strong assumption on anchor test. The anchor test should be representative in terms of test content and statistics because it is the key to disentangling the total difference confounded with test difference and group difference. If the anchor test is a lack of representativeness, it is hard to generalize the linkage function extracted from common items to the total test. The reliance on common items in NEAT design may violate the population invariance property if same examinees are exposed to new Form X and its corresponding base Form Y. If the anchor items are exposed to examinees who are going to retake the tests, these repeaters may memorize some of the items and perform better than first-time test takers on anchor test. The difference between repeater group and total group draws our attention to the problem of population variance on equating.

1.1.3 Repeater Effects on Equating

In licensure testing programs or university admission testing programs, examinee's score is always used for evaluating a candidate. To get an ideal score, examinees tend to attempt the test multiple times till they are satisfied with their final score. The examinees who take the same test repeatedly are referred to as repeaters and those who attempt the test for the first time are referred to as non-repeaters. Previous studies showed that repeater group is distinguishable from the total sample group in three situations (Andrulis, Starr, & Furst, 1978; Kim & Kolen, 2010; Yang, Bontya & Moses, 2011; Kim & Walker, 2012; Duong & von Davier, 2012; Rogers & Radwan, 2015). First, repeaters may have lower average score than the total group because they failed at the first time and the score distribution locates at the left of the total group distribution

(Duong & von Davier, 2012, Rogers & Radwan, 2015). Secondly, repeaters might have more experiences than non-repeaters and they might make substantial progress after the first administration (Kim & Kolen, 2010; Yang, Bontya & Moses, 2011). Lastly, in NEAT design with internal anchor, repeaters may have taken the previous form of the test and memorized the common items before the second administration. Thus, they may have better performance on anchor test but lower score on non-anchor than total group.

Among three situations, the first one is more likely to happen, that is, the repeaters are likely to have lower ability than total group. Including a large number of repeaters in new form can drop the observed total score and thereby making the new test form appeared harder (Andrulis, Starr, & Furst, 1978). Puhan (2009) suggested two solutions to dealing with repeater effects. He recommended either excluding the repeaters before equating or retaining all examinees but omitting the problematic items that are frequently exposed to repeaters in anchor test. The excluding repeaters solution will reduce the sample size and increase the sampling errors (Kolen and Brenan, 2014) whereas keeping the repeaters and removing common items may bias the equating function. In other words, psychometricians may face a dilemma of either keeping repeaters to ensure adequate examinees or removing repeaters to retain equity property. In addition to these two solutions, previous studies applied the equating methods under item response theory (IRT) framework because the IRT-based equating is expected to be more population invariant than observed-score equating if the assumptions of IRT model are satisfactory (Lord, 1980; Hambleton, Swaminathan, & Rogers, 1991). However, this solution would lead another problem, which is the demand for a large sample size for the use of IRT models.

1.1.4 Challenges of Small Sample Equating

Small sample equating can be performed in the small-scale testing programs such as teaching licensure tests with 20-30 examinees per test form when the new edition of the test is adopted by only a few states. To report the scores by a certain date, equating might be performed even though the sample size is very small (Kim & Livingston, 2010). Equating with a small number of examinees may cause some problems. When performing equating across multiple forms, an insufficient number of examinees may not cover the full range of score scale of the test. If the score range is larger than the number of examinees, there will be restricted range in the score distribution or sparseness of observed score on the entire score scale (Kim, von Davier & Haberman, 2008). In addition, the small sample might not be representative and thereby yielding the estimated equating function differs from that of the population (Kim, von Davier, & Haberman, 2008). Moreover, under the framework of IRT model, the equating also relies on item parameter estimation and scale transformation. The small sample causes errors in parameter estimation as well as equating function estimation, which in turns affect the scores for all examinees. The reported (equated) score with large errors would result in serious problems of validity and fairness issues.

Facing repeater problems under a context with small sample could make the situation more complicated. Excluding repeaters may not retain the group invariance and decrease the number of total examinees. It is not clear what is the minimum requirement of a sample size to decide whether to remove examinees or not. Given a certain number of examinees, if the decision is exclusion, the following up question would be which small sample equating techniques can mitigate the negative effects results from the large

decrease in examinees? Therefore, the current study is trying to answer an overarching question about how to deal with repeater effects on equating under small sample context. The specific questions are addressed in the following section.

1.2 Statement of the Problem and Research Questions

Research about small sample equating and repeater effects on equating are two areas that were studied for years, however, there are four areas that were not explored by previous research. First, few studies have combined these two topics in one study. Small equating can occur in two contexts: 1) the test is administered to a small group of examinees, 2) total group is composed of repeaters and non-repeaters, a small amount of examinees is left after removing repeaters (e.g., small-volume teaching licensure exam). Previous studies have compared different equating tools with small sample size (e.g., Livingston, 1993; Hanson, Zeng & Colton, 1994; Skaggs, 2005; Livingston & Kim, 2008; Kim, von Davier, & Haberman, 2008). However, there is a lack of literature studying the small sample equating methods if the property of group invariance did not hold due to a large proportion of repeaters. Second, in repeater effect studies, the majority of the research merely focused on excluding repeaters or not excluding repeaters, few studies had examined the other solutions such as applying IRT equating or removing problematic items that are frequently exposed to examinees. Thirdly, most of the small sample equating studies focused on the test only consists of multiple-choice items but few of them studied the impact of small samples on the mixed-format test that consists dichotomous and polytomous responses. To fit the IRT model, the mixed-format test requires larger sample size than the dichotomous model, it is necessary to study the minimum sample size on IRT equating for the test with multiple choice and construct

response items. Lastly, previous studies investigated equating bias, equating errors, the difference between non-repeater group and the total group at test level by resampling data from operational assessment. However, as an indicator of estimation accuracy, it is not clear how the large estimation errors, bias and difference at each score point impact decision making based on examinee's reported score. If the reported test score has large errors or bias after equating, the pass/fail decision might be invalid to examinees. Therefore, it is important to know the practical consequences result from group dependence or small sample size on equating.

The main purpose of the study was to compare the equating results and population invariance measures of three solutions to repeater effects: 1) excluding repeaters, 2) including repeaters but removing problematic items, 3) applying IRT Rasch equating under a context of equating with small sample size. The secondary purpose is to compare the equating methods under different test theory frameworks and investigate whether using certain equating technique can mitigate the problems of a small sample. The ultimate goal is to examine the practical implications of the results by estimating the accuracy of performance classification.

The purposes are addressed by the following questions:

1. Under the same test conditions and small sample equating techniques, how do different repeater effects solutions impact the equating results?
    i. Does the exclusion of repeater approach hold the invariance property?
    ii. Among three solutions, which one produce higher equating accuracy and lower equating bias?
2. How do different small sample equating techniques impact the equating results?

   i. Does the performance of equating techniques differ depends on test conditions and repeater effects solutions?
   ii. If there are interaction effects, which conditions produce less equating errors and bias?

3. What are the practical implications of this study?
   i. How do the equating results and population invariance affect performance classification at the individual level?
   ii. At a given condition of sample size and proportion of repeaters, what would be recommended for equating method and inclusion or exclusion approaches to get an acceptable level of equating accuracy, equating bias, population invariance, and classification accuracy

CHAPTER 2

LITERATURE REVIEW

This chapter introduces some popular equating approaches under classical test theory (CTT) and IRT frameworks, discusses the recent techniques developed for small-sample equating contexts and describes the reviews of past studies examining the repeater effects on equating.

2.1 Classical Equating

Popular classical equating methods include identity equating, the simplest equating method requiring no transformation; mean equating, score transformation only based on the mean of the distribution of scores on the new form test and old form test; linear equating, the score transformation based on the linear function between score distributions of two forms; and fourth, equipercentile equating, matching a score of Form Y that has the same percentile rank with an equivalent score on Form X.

2.1.1 Identity Equating

Identity equating is the simplest approach among all equating functions. Identity equating requires no equating transformation between the old form and the new form. In other words, identity equating is the same as no equating. The identity equating function can be formalized in equation (2.1)

$$y = ID_Y(x) = x. \tag{2.1}$$

In equation 2.1, *x* refers to the raw score on Form X and *y* refers to the equated score equivalent to *x* on Form Y. Identity equating is included here because it is commonly used in two situations: equating with extremely small size of examinees when

forms are completely parallel where the test forms have equal difficulty level and groups have equal ability level (Skaggs, 2005), or used as a baseline equating function to compare with other traditional equating relationships.

2.1.2 Mean Equating

The mean equating function is estimated based on the mean score of the old Form Y and new Form X. The equating function adjusts for the difference in mean test difficulty but maintains other statistical characteristics (e.g., standard deviation, skewness and kurtosis). The equating function is formalized in equation (2.2)

$$y = m_Y(x) = x - [\mu(X) - \mu(Y)], \quad (2.2)$$

where $m_Y(x)$ is the score x transformed to old Form Y using the mean equating function, x is a score point on the observed score scale X, $\mu(X)$ and $\mu(Y)$ are the mean score of Form X and Y for a population. The equation (2.2) illustrates the mean equating procedure. With given test score x on the new form, the corresponding score on Form Y is obtained by subtracting the difference in mean scores between two forms. In equation (2.2), the difference in means $[\mu(X) - \mu(Y)]$ is a constant that applies to each score point. In other words, the score distribution of two forms differ only in this constant. If the mean score of new Form X is 3 points lower than mean score of Form Y, three points need to be added to every score point in Form X to transform to the Form Y scale. The mean equating is simple and easy to apply but it assumes the distance between Form X and Form Y is equal along all score levels. In practice, the transformation between forms might not be constant. For example, the mean score indicates Form X is 3 points easier than Form Y at high proficiency level but low proficiency examinees are likely to obtain similar scores on two test forms. Therefore, the equating relationship requires a mean

score transformation less than 3 points for examinees with scores at the lower end of the score scale.

2.1.3 Linear Equating

Unlike mean equating where the differences between two forms are equal along score scale, linear equating is performed based on a linear conversion from Form X to Form Y. The function adjusts the mean and standard deviation by standardizing the Form X and Form Y on the same scale (i.e., z-score scale). The linear conversion equating is

$$y = l_Y(x) = \frac{\sigma(Y)}{\sigma(X)} * x + \left[\mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\mu(X)\right], \quad (2.3)$$

where $l_Y(x)$ is the linear equating function for score x on scale Y, $\sigma(Y)$ and $\sigma(X)$ are the standard deviations of Form Y and Form X, respectively.

When the standard deviation of the score is 1 for both forms, the equation (2.3) is equivalent to equation (2.2). That is, the mean equating function can be considered as a special case of linear equating function. Although the linear equating function in equation (2.3) provides more flexibility along the score scale, it has some limitations. Firstly, it is possible that the range of equated score is beyond the score range. On a 0-100 score scale, the highest equated score can exceed 100 if the old form has a lower difficulty level than new Form X. One common solution is to truncate the equated score so that the highest above 100 is equal to 100 and the equated score below 0 is constrained to 0. The other limitation is related to the assumptions. The use of linear equating function assumes the relationship between two forms is linear and score distributions of two forms are identical. To accommodate the test conditions where these prerequires are not

satisfactory, equipercentile equating, as a nonlinear equating procedure, was developed by Braun and Holland (1982).

2.1.4 Equipercentile Equating

The equipercentile equating function transforms the raw score of Form X to Form Y by using a nonlinear conversion, the equating process is performed based on the assumption that scores on Form X and Form Y are continuous random variables. However, the test scores, especially number-correct observed scores are discrete variables. Therefore, the equating process is performed by matching the percentile rank of each discrete score. The process of equipercentile equating is defined as

$$y = e_Y(x) = G^{-1}[F(x)], \tag{2.4}$$

where $e_Y(x)$ represents the equipercentile function that rescales Form X score on Form Y, G is the cumulative function of Form Y, $G^{-1}$is the inverse of the cumulative function, $F(x)$ is the cumulative function of Form X. The equation (2.4) describes a 3-step process: specifying the percentile rank of score x on form X, estimating the corresponding percentile in the Form Y and then matching the equipercentile equivalent score x on the Form Y with the same percentile rank.

Unlike mean equating and linear equating, equipercentile equating can limit the equated score within the possible score range. However, the equipercentile equating function requires more estimated parameters. Given the same sample size, the equipercentile approach might produce less precise results due to sampling errors. If the equipercentile equating is applied with deficient sample size, there might be sparseness at certain score points and leads to irregularity between score distribution and the equipercentile relationship. As a result, smoothing techniques are widely used in previous

research and practice under random group design and non-equivalent group design (Kolen, Brennen, 2014). Among all smoothing techniques, log-linear presmoothing was implemented to equipercentile equating in most small equating studies (e.g., Livingston, 1993; Han, Zhang, & Colton; 1994). The log-linear presmoothing is performed on raw score distribution before equating procedure. This approach estimates multiple polynomial models that fit the log of the density function of the raw score. These multiple polynomial functions differ in degree of polynomial *C*. The degree of each polynomial term determines raw score distribution that is preserved in the smoothed distribution (Holland & Thayer, 1987). If the fitted distribution preserved the first three moments (i.e., I =3), then the mean, variance, skewness of observed distribution are preserved. Typically, the choice of the final polynomial function depends on the resulting random error, systematic error, overall accuracy and improvement in model fitness. For more detailed explanations of pre-smoothing and post-smoothing techniques under different data collection designs, see Kolen and Brennen (2014).

2.2 IRT-Based Equating

In addition to classical equating based on observed scores, many testing programs use item response theory (IRT) models to assemble tests and IRT equating to adjust the scores on the same scale. Kolen and Brennan (2014) described a three-step procedure to perform IRT-based equating for unidimensional IRT models: 1) estimate the item parameters and ability parameters by fitting the data with an appropriate IRT model, 2) transform estimated item parameters of an alternative form onto the base form scale, 3) establish a raw-to-raw/true score conversion table by using IRT observed score equating

(OSE) or true score equating (TSE). This section reviews popular unidimensional IRT models and the procedures of IRT TSE.

2.2.1 IRT Models

The most popular IRT models for dichotomous data are one-parameter (Rasch), two-parameter logistic (2PL) model, and three-parameter logistic (3PL) model. The Rasch model, which is also referred to as the one-parameter logistic (1PL) model, was introduced by Rasch (1960). In the Rasch model, item difficulty is the only item parameter that determines the response patterns with given proficiency level. Birnbaurm (1968) introduced the 2PL model with a discrimination parameter representing the slope of the *Item Characteristic Curve* (ICC).

The Rasch model and 2PL IRT model are written as

$$P\left(x_{ij}=1\middle|\theta_j,b_i\right)=\frac{\exp[D(\theta_j-b_i)]}{1+\exp[D(\theta_j-b_j)]} \tag{2.5}$$

and

$$P\left(x_{ij}=1\middle|\theta_j,a_i,b_i\right)=\frac{\exp[Da_i(\theta_j-b_i)]}{1+\exp[Da_i(\theta_j-b_j)]} \tag{2.6}$$

Where θ_j is the true ability for examinee j, b_i represents the difficulty parameter, a_i is the discrimination parameter. D is a scaling factor to make the logistic model close to normal ogive function. The 3PL model is the extension of the 1PL and 2PL models. This model has a pseudo-guessing parameter c_i that denotes the probability of less capable examinees answering the item correctly by guessing. The model is expressed as

$$P\left(x_{ij}=1\middle|\theta_j,a_i,b_i,c_i\right)=c_i+(1-c_i)\frac{\exp[Da_i(\theta_j-b_i)]}{1+\exp[Da_i(\theta_j-b_j)]}. \tag{2.7}$$

Apparently, 2PL and Rasch model can be considered as a special case of the 3PL model. Rasch model is nested within the 2PL model, and the 2PL model is nested within 3PL model.

2.2.2 Linking and Scale Transformation

When performing equating with the NEAT design using an IRT model, the parameter estimates of items need to be placed on the same scale using linear transformation equations if the IRT model holds. The linear relationship between item parameters on the two scales X and Y is defined as

$$a_Y = \frac{a_X}{A} \tag{2.8}$$

$$b_Y = A(b_X) + B \tag{2.9}$$

$$c_Y = c_X. \tag{2.10}$$

In the above equations, *A* and *B* are constants in the linear equation. The item parameters a_Y, b_Y and c_Y are parameters of the base form Y while parameters a_X, b_X and c_X are parameters on the new form X. The θ-values for examinee *i* for the two scaled are related as follows:

$$\theta_{Yi} = A(\theta_{Xi}) + B. \tag{2.11}$$

There are different ways to estimate *A-constant* and *B-constant*, the widely-used methods are the moment methods (the mean/mean and mean/sigma; Loyd & Hoover, 1980; Macro, 1977) and the test characteristic curve (TCC) methods (Haebara, 1980; Stocking & Lord, 1983). The mean/mean method uses the mean of the *a*-parameters and the mean of the *b*-parameters of common items to estimate *A*-constant and *B*-constant whereas the mean/sigma method estimate linking constants only based on the means and

standard deviations of the *b*-parameter estimates from common items. Unlike moment methods, the TCC method considers all parameter of anchor items on two scales simultaneously. Haebara approach (1980) and Stocking and Lord approach (1983) are very similar. Haebara approach estimates linking constants based on the sum of the squared difference between item characteristic curves (ICCs) for given θ over common items while Stocking and Lord approach uses the squared difference of TCCs of anchor test for given θ. Many studies have compared the moment methods and characteristic curve methods. It is found that the TCC methods provided more stable results than moment methods (*Baker & Al-Karni, 1991*; Hanson & Beguin, 2002; Ogasawara, 2001).

2.2.3 True Score Equating

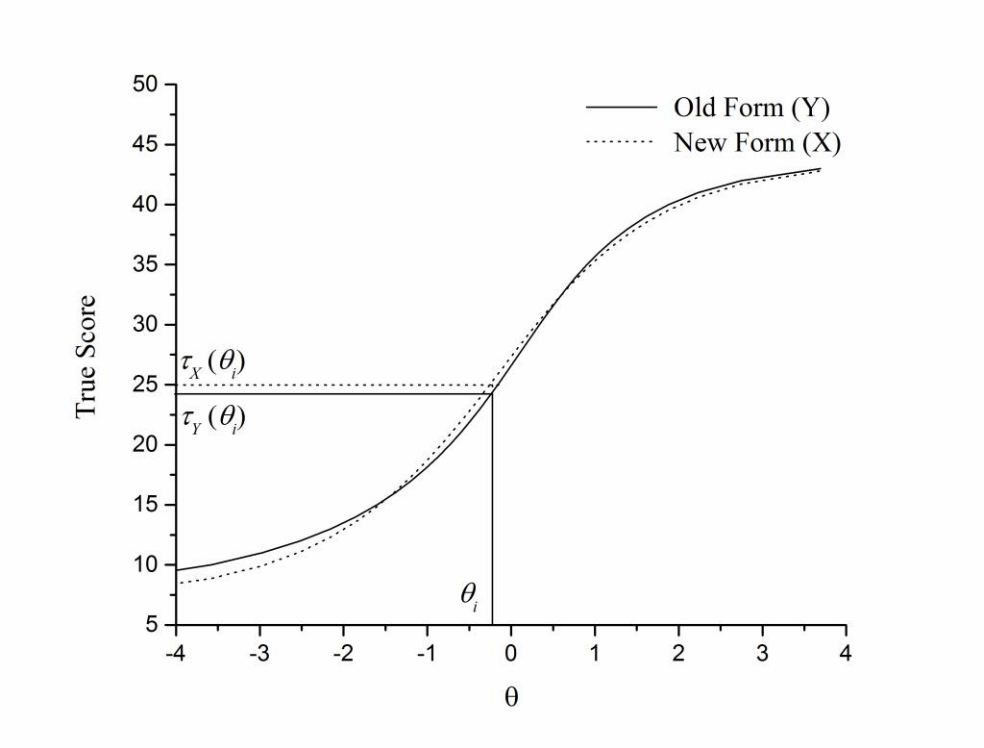
After placing the item parameters on the same scale, the next step is to relate number-corrected scores on Form X and Form Y. This process can be conducted using two methods: IRT observed score method and IRT true score equating. The OSE produces an estimated distribution of observed number-correct scores on each form and then performs the equating using equipercentile methods. The TSE established a conversion table where the true score of one form $\tau_X(\theta_i)$ associated with a given θ is considered to be equivalent to the true score $\tau_Y(\theta_i)$ of another form related with the same θ. The process is denoted as

$$irt_Y(\tau_X) = \tau_Y(\tau_X^{-1}), \tag{2.12}$$

where τ_X^{-1} is the θ_i that is corresponding to the true score τ_X on Form X, τ_Y is the corresponding true score on Form Y and $irt_Y(\tau_X)$ is an integer score of τ_X. Equation (2.12) describes a three-step procedure that is presented in Figure 2.1. First, a true number-correct score τ_X is specified and the true score is typically an integer on Form X

scale. In Figure 2.1, the integer score is 25. The next step is to find the θ_i that is corresponding to $irt_Y(\tau_X)$ through TCC on the new form. The last step is to find the true score τ_Y, which corresponds to same θ_i, via mapping from TCC of old Form Y.

Figure 2.1. TCCs of old Form Y and new Form X



Although IRT TSE and IRT OSE differ in the computation process, previous research showed they produce similar results for the 3PL model in NEAT design (Lord and Wingersky, 1984; Kolen and Brennan, 2014). For Rasch equating, true score equating is more commonly used in practice (Kolen and Brennan, 2014).

2.3 Small Sample Equating

Small sample equating was designed to be performed in the small-scale testing programs. For example, a small sample equating situation occurs in teaching licensure tests with 20-30 examinees per test form when the new edition of the test is adopted by

only a few states. To report the scores by a specified date, equating might be performed even though the sample size is very small (Kim & Livingston, 2010).

There is no general consensus of what size would be the lower limit of a "small" sample size. Kolen and Brennan (2014) recommend a minimum size of 400 per form for linear equating and at least 1,500 to perform equipercentile equating. Among studies focusing on small-volume testing programs, sample sizes range from 10 (Kim, von Davier, & Haberman, 2008) to 3000 (Hanson, Zeng, &Colton, 1994) have been studied with the random group design or non-equivalent group design. This section introduces and reviews studies using conventional small-sample equating approaches and new methods that were developed since 2008. The most conventional approaches for small sample equating situation was to apply smoothed techniques to equipercentile equating (e.g., Livingston, 1993; Hanson et al, 1994; Skaggas, 1995). The new methods include circle-arc (Livingston & Kim, 2008; Livingston & Kim, 2009); empirical Bayes procedure using collateral information (Kim, Livingston, & Lewis, 2011); synthetic linking function (Kim & Livingston, 2010; Kim, von Davier, & Haberman, 2011); nominal weights mean equating (Babcock, Albano, & Raymond, 2012); and general linear function (Albano, 2015). Under the IRT framework, Rasch model is more appropriate than multiple-parameter IRT model for the small size of examinees (Kolen and Brennen, 2014).

2.3.1 Traditional Small Sample Equating

Livingston (1993) compared the log-linear presmoothing with no smoothing using chained equipercentile equating with samples of 25, 50, 100 and 200 examinees in a random group design. The data were resampled from Advanced Placement History

Examination. The results showed that presmoothing significantly improved the overall accuracy based on the RMSD. However, the results revealed that the limitation of log-linear smoothing was that increasing number of moments reduced random error but introduced more systematic error, especially if more than four moments were preserved. In other words, this technique could offset the decrease of the standard error of equating but increase the bias of equating.

Hanson et al. (1994) compared identity equating, linear equating, and unsmoothed, presmoothed and postsmoothed equipercentile equating in the random group design for five ACT assessment tests in which 100, 250, 500, 1000, and 3000 examinees were resampled. The results indicated that identity equating produced less equating errors than other equating methods with a sample size of 100. In addition, applying presmoothed and postsmoothed equipercentile equating could significantly decrease the sampling errors, and the difference between pre- or postsmoothing was negligible. Hanson et al. (1994) suggested the minimum sample size for smoothed equipercentile equating is 250.

Parshall, Houghton, and Kromrey (1995) examined the standard errors and bias of equating with linear equating in the NEAT design. The data sets were resampled from a state teacher certified test with the size of 15, 25, 50 and 100. Three important findings were drawn from this study. The first finding was that standard error substantially increased as the sample size decreased. The overall error increased sharply under extremely small sample size conditions (N =15). Secondly, the sampling error was smallest at the middle score points and there was a clear pattern that the errors monotonically increased as the scores deviated away from the mean. Similar to standard

errors, the values of bias are smaller at the points closer to the mean score. The last finding was that the test with the highest correlation between the anchor and total test produced the least amount of equating errors.

Skaggs (2005) compared identity equating, mean equating, linear equating, unsmoothed and log-linear presmoothing equipercentile equating with sample sizes of 25, 50, 75, 100, 150 and 200 in random group design. The data were resampled from the Tests for General Educational Development (GED) over 110,000 examinees. The findings of Skaggs (2005) study were similar to Hanson et al. (1994) and Parshall et al. (2005). The results showed that standard errors decreased as the sample sized increased from 25 to 200. When the sample size was smaller than 50, identity equating was preferable than other equating approaches. Log-linear presmoothing technique could significantly reduce the sampling errors, however, the improvement in equating accuracy was trivial if the log-linear models fitted to smoothing beyond three moments.

2.3.2 Circle-Arc Equating

Livingston and Kim (2008) proposed two circle-arc equating approaches (systematic approach and simplified approach). The estimated equating function is an "arc curve" of a circle. The circle is determined by passing through a lower end-point, one higher point and a middle point. The lower and higher points are the lowest and highest points of each form. Under the NEAT design, the middle point is the equated mean score from new to reference form. The systematic circle-arc equating is a curvilinear equating function while the simplified function is composed of a linear component and curvilinear component. although two approaches have different formats, it was found that these two approaches had very similar results. Livingston and Kim

(2009) compared circle-arc equating with identity equating, mean equating, chained linear equating and log-linear presmoothed chained equipercentile equating where 25 examinees took the new form and 75 examinees took the old form under a NEAT design. The examinees took each form had the same proficiency level but the values of effect size between the average test scores were substantially different. The results showed that circle-arc methods had the smallest overall RMSD than mean equating and chained equipercentile. In terms of the equating bias, circle-arc equating produced more bias in the middle of score distribution while linear equating and equipercentile equating produced more bias at the two ends of the score scale.

2.3.3 Synthetic Linking Function

Kim et al. (2008) introduced a synthetic linking function that combined identity function and a traditional equating function (e.g., chained linear function) under the NEAT design. The weight of each function is ranged between 0 to 1. The sum of two weights should be equal to 1. They compared the synthetic function with identity function and chained linear function for the size of 10, 25, 50, 100 and 200. One data set had a significant difference between means of the anchor tests while the other had negligible difference. Across all pairs of data set, two forms had similar test difficulty levels for both overall tests. The results of this study showed that the identity function produced the smallest error but large bias when the sample size was smaller than 50. The chained linear method showed the smallest bias but greatest amount of errors over all sample sizes. One finding of this study indicated that synthetic linking function exhibited relatively low equating error at the expense of a large amount of bias. Another finding suggested that synthetic function might be an alternative to identity equating if the groups

differed in ability. Kim, von Davier and Haberma (2011) conducted a followed-up study investigated the performance under a non-parallel situation. The research design was similar to that of the Kim et al. (2008) study except for the substantial difference in test difficulty between new and old forms. The results of Kim et al. (2011) showed that the identity equating was most appropriate when the sample size is smaller than 25. When the sample size is larger than 25, the chained linear equating slightly outperformed synthetic linking function, followed by identity function. This study concluded that synthetic function might not be an appropriate choice if new and old forms had significantly different test difficulty level.

2.3.4 Nominal Weights Mean Equating

Babcock, Albano, and Raymond (2012) introduced a new equating method under NEAT design, which is referred to as the nominal weight mean equating. The nominal weight mean equating is a simplified version of Tucker linear equating where the term of covariance and variance is replaced by the number of items of the total test to anchor test. For more details about Tucker Method and other methods for NEAT design, see Kolen and Brennen (2014). Unlike other small sample equating studies where the response data were resampled from real data sets, Babcock et al. (2012) simulated response data based on 3PL IRT models with sample size 20, 50, and 80. The mean difference in test difficulty was categorized into three levels: $(b_A - b_B) = 0$, $(b_A - b_B) = 0.35$ and $(b_A - b_B) = 0.70$. The study consists of three sub-studies differed in simulation conditions. In study 1, examinees took parallel forms had the same ability distribution. In study 2, examinee groups took Form A (hereafter referred as GA) had lower ability level than examinee group took Form B (hereafter referred as GB). In study 3, GA had higher proficiency

level than GB. Five equating methods, mean equating, nominal weight mean equating, identity equating, synthetic equating function and circle-arc equating were conducted across all conditions. The results showed that for sample size smaller than 50, identity equating was the most accurate method when the test forms were equally difficult. However, the identity function produced larger bias as the difference in test difficulty increased. Compared to synthetic equating and identity equating, nominal weight mean equating and circle-arc equating were more tolerable to group difference and nonparallel forms. When the sample size was equal to 50 and 80, nominal weight mean equating produced the least amount of bias and errors. As a result, the authors suggested nominal weight mean equating as a promising alternative equating method for small-sample equating.

2.3.5 Empirical Bayes (EB)

Kim, Livingston, and Lewis (2011) investigated the improvement of equating accuracy by incorporating collateral information from prior equating procedures, this procedure is referred as empirical bayes (EB) procedure. The equating function estimated by the EB procedure is a product from current equating and prior equatings. Therefore, this procedure does not require a large sample size because the estimated equating results partially rely on the collateral information from prior equating results. They compared the equating results obtained from non-EB procedure and EB procedure using chained linear, chained mean and synthetic linking equating approaches. The other factors involved in the study are sample size (N = 10, 25, 50, 75, 100, and 200) and number of prior equatings (0, 3, 6, 9 ,12). Generally speaking, the EB procedure produced smaller amount of bias and less equating errors than non-EB procedure. However, when the current pair

of forms was not sampled from the same pool as the pairs of forms in the prior equatings, the EB procedure yielded larger equating errors than non-EB procedure. Nevertheless, if the current equating and prior equatings were from the same equating pool, the overall equating errors decreased as the number of prior equating increased. In terms of the comparison among three equating approaches, chained linear equating produced the lowest RMSE across all conditions regardless of EB or non-EB procedure.

2.3.6 General Linear Equating

Albano (2015) introduced another method for small sample equating problem. Similar to synthetic linking function, the new approach is a general linear function that is presented as a general form of mean, linear, and identity function. Albano (2015) compared the general linear function with identity, mean, linear, circle-arc, synthetic, and equipercentile equating methods with sample size 20, 50, 100, and 500 in a single-group design. To examine the performance of each equating method under the situation with unbalanced proficiency levels between test forms, the mean score of reference form was artificially adjusted downward from 160.7 to 148.0, which is 10.5 points lower than the new form. The results showed that general linear equating produced the smallest amount of root mean squared error (RMSE) across all levels of sample size of 20, 50, and 100. Equipercentile equating gave the lowest RMSE when $N = 500$.

2.3.7 Summary of Small Sample Equating

Research on small sample equating have been studied since early 90s. Among all small sample equating techniques, identity equating, mean equating, linear equating, presmoothing and postsmoothing were the most traditional approach used in practice and

research. In previous studies, the presmoothing equipercentile equating was frequently used as a criterion equating to compare the performance of other new developed equating methods (e.g., Livingston & Kim, 2009; Kim and Livingston, 2010). Under NEAT design, applying presmoothing procedures on chained equipercentile equating significantly reduced equating errors when the sample size is moderate (Hanson et al., 1994; Skaggas, 2005). However, previous studies showed that this approach did not perform well when the sample size is smaller than 50. In addition, the equipercentile equating was likely to produced greater standard errors at two ends of score scale than other equating approaches (Livingston & Kim, 2009). As a result, presmoothing equipercentile equating might not be the best choice for a testing program with extremely small sample size (e.g., $N < 50$).

The general linear equating function was a newly developed method that is very flexible in form and can be generalized to multiple forms by manipulating the weight combination of each component (i.e., the weight of identity function, mean function, and linear function). Albano (2015) found that this method produced smaller amount equating errors and bias than circle-arc and synthetic equating regardless of the difference in test difficulty between pairs of the form. This method might be promising for equating with different levels of sample size. However, the main drawback of the method is that little research has explored the use of the general linear function under different test conditions. That is, there is no guidance to show how to manipulate the weight of each equating function component given certain levels of sample size, group difference, and the difference in test difficulty. The lack of literature might cause cumbersome in applying general linear equating function in the current study. Similarly, the synthetic

linking function also lacks the detailed instruction in deciding the weight of identity function and traditional equating function. Kim, von Daver, and Haberman (2011) found that unequal weight for identity and traditional equating function outperformed equal weight. Unfortunately, the study did not fully explain why the unequal weight was preferred and how to choose the weight or decide weight for other data sets. The absence of literature is the main issue for newly developed equating approaches. To use the EB procedure, more studies need to be conducted to show how to incorporate collateral information into current equating procedures. Although their study found EB procedure had an outstanding performance when the current equating and prior equatings were from the same equating pool. They did not answer the questions that how to detect whether the prior and current equating were from the same equating population, and how to form equating pool. Therefore, these equating methodologies are not considered for current study because of the absence of detailed information.

Under the NEAT design, circle-arc equating and nominal weight equating might be easier to apply in practice than other newly proposed methods. Unlike general linear equating and synthetic linking function, the formats of the circle-arc equating and nominal weight equating are fixed because there is no need to decide the weight of each component of the equating function. In addition, these two methods perform better than identity equating and synthetic equating under the situations with unequal group ability distribution and test form difficulty. Lastly, these two methods had the most stable performance regardless of score difference across all levels of sample size (Bacock et al., 2012). In practice, there is always a possibility that the observed score collected from new and reference forms are substantially different. However, it is hard to control or

eliminate the difference for small-volume testing programs because these programs lack sources such as item statistics or item pools. As a result, circle-arc or nominal weight might be promising solutions because they have a low requirement for equivalent forms, group difference, prior information, and sufficient sample size.

2.3.8 Rasch True Score Equating

The use of Rasch equating requires a smaller sample size than 2PL and 3PL IRT models because of the feature of simplicity (Kolen & Brennen, 2014). To obtain stable parameter estimates, parameter recovery studies suggested the minimum sample size for Rasch modeling was 100 (Stone, Yumoto, & Dale, 2003; Chen et al., 2013). Linacre (1994) proposed that the minimum sample size for dichotomous and polytomous response data were 30 and 50 to get an acceptable estimation. Although Rasch modeling requires a relatively small sample size, Rasch equating was not mainly developed for small-sample equating as the classical equating approaches that are reviewed in the previous section. In fact, the use of Rasch equating heavily relies on strong assumptions such as unidimensionality and model-fit. According to Linacre (1994), the minimum sample size may differ depends on test length, the purpose of the test, homogeneity of the population, and other factors. The best way to discover the appropriate sample size for Rasch model on certain data sets is to conduct simulation study and examined the estimation accuracy across different sample size conditions.

2.4 Repeater Effects on Equating

Examinees who take the test the repeatedly are frequently observed in educational assessments and certification testing programs. Previous research reported that the

percentage of examinees that retook National Council of the State Boards of Nursing Licensure Examination (NCLEX) was approximately 20% (Gorham & Botempo, 1996). Repeaters constituted 40% of the total sample from several test sites for Swedish Scholastic Assessment (SweSAT; Stage and Ögren, 2004). Thornton, Stilwell and Reese (2006) reported that percentage of repeaters took the Law School Admission Test (LSAT) was around 22% between 2001 and 2005. A recent study showed that approximately 18% and 16% repeaters took the Test of English for International Communication (TOEIC) in Asian countries for speaking and writing, respectively (Liao and Qu, 2010).

Previous repeater effects studies focused on how the inclusion of repeaters influenced ability estimates and reported scale scores, a limited number of studies concerned the repeaters effects on equating. For tests requiring equating procedures, the reported scores are derived from equating function, if the inclusion of repeaters has an impact on equating, consequently, it would impact the scale scores as well as the decision made based on the equated score. Andrulis, Sttar, and Furst (1978) investigated the effects of repeaters on linear equating with a random group design where 273 examinees took the old form and 172 examinees took the new form. Among 172 examinees, 20 of them were repeaters. They found that including repeaters, whose performance was lower than the total group, strongly influenced the linear equating function by "lifting up" the equated score of the new form and make an additional 3% of examinees passed the test. Though Andrulis et al. (1978) made a suggestion of removing repeaters before equating and then applied the equating conversion to all examinees, recent studies had mixed results regarding repeaters effects on equating.

Table 2.1 provides a summary regarding equating methods, equating designs, the evaluation criteria, and conclusions that made based on the results. It was clear that studies conducted since 2009 included more evaluation criteria rather than the difference between average test scores; moreover, the conclusions are more complicated than routinely removing repeaters before equating.

Table 2.1. Summary of Repeaters Effects on Equating Studies

Study	Equating Method	Equating Design	Criterion	Conclusion
Andrulis, Starr, Furst (1978)	Linear Equating	Random	mean score of the total test	Exclude repeaters
Puhan (2009)	Chained linear Chained Equipercentile	NEAT	conditional difference curve (CDC) root expected squared difference (RESD)	Exclude Repeaters
Kim and Kolen (2010)	Equipercentile 3PL IRT concurrent calibration	Random	RMSD REMSD equally weighted REMSD. (ewREMSD) standardized root squared difference (RSD) Root weighted-average squared difference (RWSD)	Exclude repeaters
Puhan (2011)	Chained linear CL and pre-smoothing Chained equipercentile	NEAT	CDC, RESD	Include under small-sample size

Yang, Bontya, Moses (2011)	Smoothed chained equipercentile Chained linear	NEAT	RESD RMSD	Exclude repeaters
Kim & Walker (2012)	Pre-smoothed Chained equipercentile	NEAT	RMSD REMSD ewREMSD	Classify repeaters and only exclude reference repeaters
Rogers & Radwan (2015)	Forward fixed common-item parameter Modified 1PL model with fixed guessing parameter	NEAT	RMSD	Include repeaters

2.4.1 Equating Design and Equating Methods

Table 2.1 shows that 5 out of 7 studies were conducted using a NEAT design while 2 studies were performed under a random groups design. The summary reflects the prevalence use of the NEAT design. Unlike random group design, the NEAT design does not assume the examinees administered to different forms were from equivalent groups. Table 2.1 lists that chained linear and chained equipercentile equating were the most commonly used classical equating methods in non-equivalent groups design. Compared to classical equating, IRT-based equating was less commonly used based on the summary. Another observation regarding the equating method is that a smoothing technique was not required with a large sample size (N>2000). Studies with a smaller number of examinees (N < 1500) were likely to apply presmoothing equating approaches to prevent the irregularity of the score distribution (e.g., Puhan, 2011; Yang et al., 2011).

2.4.2 Evaluation Criteria

Most of the evaluation criteria listed in Table 2.1 were referred to as equitability indices and developed for checking the invariance property of equating function (Dorans and Holland in 2000). As it is introduced in the previous chapter, the invariance property of equating holds if the equating function is identical to all subgroups. In the current study, the total group is comprised of repeaters and non-repeaters.

According to the summary in the table, the most widely used measures to evaluate the degree of equating variance were conditional RMSD, a measure evaluating the equating difference between a specific subgroup and total group at each score level by taking account the proportion of examinees at the subgroups; root expected mean square difference (REMSD), an index representing the summation of RMSD values by accounting for the proportion of examinees at each score point. Kim and Kolen (2010) compared the performance of equally weighted REMSD (ewREMSD), an index used the same weight over score points, the results show that ewREMSD and REMSD had very similar values. In addition to the criteria representing the equating difference between subgroups and total group, the difference that matters (DTM) was a baseline to evaluate whether the difference is too large to cause a concern. The DTM is the half unit of the reported score if the invariance measures exceed DTM at cut-score points, it indicates the violation of invariance may have practical significance to equating. Some of the studies also depicted CDC along with the score scale. The CDC indicates the difference between total equating function and subgroup equating function, however, CDC does not take account the weight of group membership and can be negative or positive. The size of the CDC can be evaluated by DTM as well.

For credentialing tests, it is more important to investigate the population invariance measures at score point that decides pass/fail decisions. If the values of the population invariance criteria at the two-end of score points beyond the values of DTM but the value of criterion at a cut-score point within the DTM, the violation of invariance may not have a strong practical significance to a decision.

2.4.3 Review of Findings

The results varied depending on sample size, group memberships, and equating design and evaluation criteria. One early study found that including repeaters in the equating procedure favored the less capable examinees and can drop the cut scores on the new form test. Thus, the equating function derived from repeaters was biased and inaccurate (Andrulis et al., 1978). However, recent studies had mixed findings regarding repeaters effects.

Puhan (2009) compared the effects of inclusion and exclusion of repeaters on equating results in a NEAT design using chained linear and chained equipercentile equating. The data were collected using two large-scale credentialing assessments, Test A and Test B, with non-repeaters in the old form and total sample (i.e., repeaters and non-repeaters) in the new form. For Test A, there were 580 examinees taking old form and 1117 examinee taking new form with 80 repeaters. For Test B, there were 534 and 1110 people taking old and new form with 164 repeaters in the new form. For both tests, the new form sample appeared less able (i.e., lower mean score on the anchor test) than old form sample if the repeaters were included. The measures evaluating the equating difference between total group and non-repeaters were CDC and RESD, these two measures were evaluated by DTM to show the degree of equating variance. The results

showed for both tests, excluding repeaters had little impact on properties of invariance of equating regardless of the equating methods.

To examine the repeater effects on passing rates, Puhan (2011) used the same set of data, data collection design and equating methods investigated the passing rates at score points that corresponded to the 10^{th}, 25^{th}, 50^{th}, 75^{th} and 90^{th} percentile. The results showed that passing rates were identical for high percentile ranks between non-repeaters and total sample group. However, the passing rates at the low percentile points derived from non-repeaters are slightly lower than the passing rates obtained from total group. Next, the author resampled 100 and 50 examinees as two sample size conditions where 30% of the total sample were repeaters. When the sample size was 50, in which15 of them were repeaters, the random equating error became larger than DTM at the 10^{th} and 25^{th} percentile points. As a result, the study suggested keeping repeaters when the sample size was small.

In Kim and Walker (2012), the total sample was categorized into three groups: 1) non-repeaters, 2) repeaters who took the exact reference form at the previous administration (hereafter referred to as reference repeaters), 3) repeaters who took any form other than the reference form (hereafter referred to as non-reference repeaters). Data sets were collected from two large-scale licensure tests and equated using presmoothing equipercentile equating and chained linear equating. Two linear regression models were performed predicting non-anchor score from anchor item score, repeater membership (reference vs. non-reference), and interaction of membership and anchor score. The regression results showed when holding the score of anchor tests constant, reference repeaters had significantly lower overall performance on non-anchor items than non-

reference repeaters. Next, the authors computed equating difference measures between non-reference repeaters and total groups and equating difference between reference repeater and total group (i.e., RMSD, REMSD and ewREMSD). The results showed that, equating differences were negligible between non-reference repeaters and total group. However, if the repeaters were referred as the examinees who have been exposed to the common anchor items, equating function was substantially different from total group across all score levels. The results of this study indicated the importance of specifying the membership of repeater groups because reference repeaters may be advantaged by item exposure. At last, the authors suggested a solution that the equating can be performed by only excluding reference repeaters.

Roger and Radwan (2015) examined the effect of repeaters on equating and passing rates by manipulating the percentage of repeaters. The data were collected from a large-scale literacy test for English learners from a matrix-sampled design. In this design, the items in new operational form were field-tested items in old operational forms. There were 24 different sets of field test items embedded within the operational forms of the previous year. The purpose of using matrix sampling design was to prevent repeaters remembering all common items in the previous year. From the total data set (approximately 150,000 examinees), eight pairs of equating samples were created with different percentage of repeaters in the new form, ranging from 5%-40% by decreasing the number of non-repeaters. For each pair of equating sample, the old form sample only included non-repeaters while new form had both repeaters and non-repeaters. The forward-fixed common-item parameter (FCIP) procedures for nonequivalent groups were performed by fitting the data with modified one-parameter IRT model with fixed

guessing parameter c = 0.2. The values of RMSD showed that the difference between non-repeaters and the total group became larger as the percentage of repeaters increased. In addition, the passing rates for total group were more similar to the passing rates of the population than passing rates for non-repeaters only group. As a result, the authors suggested including repeaters if anchor tests were not exposed and the proportion of repeaters was smaller than 20%. By including the repeaters, total sample was more representative of the population with respect to passing rates.

Kim and Kolen (2010) examined the repeater effects under a random group design using equipercentile equating and 3PL concurrent calibration. The results were similar to that of previous studies: 1) repeaters were likely to have less able performance than the total group, 2) excluding repeaters did not significantly impact the population invariance or passing rates. In addition, the study examined the equating difference between repeaters and non-repeaters by computing conditional RSD across score levels. It was found that the values of RSM were likely to exceed DTM at extreme lower and upper score scale. Moreover, this study compared the population invariance criteria between classical equating and IRT TSE. The results showed that equipercentile equating tended to produce larger error statistics than IRT TSE. Furthermore, the error statistics yielded by equipercentile equating were more likely to exceed the DTM than IRT TSE. These results may suggest a third solution to mitigate the population dependence due to repeater effects – applying IRT-based TSE equating.

2.5 Conclusion

According to reviews of the literature, most of the studies focused on repeater effects to large-scale assessment rather than small volume assessment, it is not clear what

are the consequences of equating and passing rates if a large proportion of repeaters are removed from a small sample size. On the other hand, retaining the repeaters may also produce a large amount of bias because repeaters are likely to have lower proficiency level and drop down the observed score of the total group. As a result, the total scores of the old form group and the new form group are not equal. According to the previous studies, three solutions were suggested for tests with a certain proportion of repeaters: removing repeaters, retaining all repeaters but removing all problematic anchor items, applying IRT equating to retain the invariance property. The first two solutions involved classical small-sample equating approaches that are reviewed in the previous section. The circle-arc equating and nominal weight mean equating are chosen because they are easy to apply, and they perform well when forms have unequal difficulty or groups with unequal ability. The nominal weight mean equating is a simplified equating approach for Tucker linear equating where the ratio of covariance and variance is replaced by the ratio of test lengths. The simplicity makes the nominal weight mean equating demands a fewer number of estimated parameters and is less susceptible to errors if the variance and covariance of the anchor or total test are not well estimated. In other words, the nominal weight equating might be a promising solution for small-sample volume tests in which the anchor items were memorized by repeaters because the variance of the anchor is ignored. The last solution for repeater effects is applying IRT equating. Under the current context, the Rasch model is more appropriate than other IRT models because it only estimates one item parameter and requires smaller sample size than other models.

Besides the equating methods that were used for previous studies, one observation of methods section reveals that most studies were conducted based real data

or resampled from real data. However, the current study simulates the response data because it is easier to control the manipulating factors and exclude other confounding factors that are not relevant to the study. Five crossed factors considered in this study are the number of examinees, the percentage of repeaters, repeater ability, equating methods, and presence of problematic anchor items that favor repeaters. The repeaters were simulated as the test retakers who are less able than total sample, have taken the exact reference form, and have seen the common items before taking new form (reference repeaters). The final goal of this study is to provide the performance of three solutions to small-volume testing programs with given sample size and proportion of repeaters. More details regarding data simulation and research methods are fully described in Chapter III.

CHAPTER 3

METHOD

3.1 Methods Overview

This chapter describes the design to investigate three solutions to mitigate errors and bias resulted from small-sample equating with a relatively large proportion of repeaters under NEAT design with an internal anchor. Three solutions were: (1) excluding repeaters, (2) removing problematic anchor items, (3) including repeaters and all anchor items but applying IRT TSE. Three solutions were compared based on equating invariance index, equating bias, equating accuracy, and decision accuracy (DA). The secondary purpose was to compare different small-sample equating approaches: (1) circle-arc equating, (2) nominal weight mean equating, (3) identity, (4) Rasch TSE.

There are two benefits of using simulation procedure for the current study. First of all, the generated data were simulated based on given true item parameters and ability parameters. As a result, it is easy to compare the estimated equating results with true equating results that were derived from true parameters. The other benefit is that simulation study can simulate some extreme conditions. In addition, the simulation study can avoid extraneous or confounding factors that arise from the real data but not relevant to the current study. Two groups of examinees were randomly selected from two populations, one took the old form while the other took the new form. To create a "worst-case scenario", it was assumed that repeaters took the new form reviewed the anchor items in the old form and even memorized some anchor items and then took the same set of anchor items in the new form. It was likely that repeaters perform better on anchor test than non-anchor test although their overall observed score was still lower than non-

repeaters. As a result, the current study simulated a context that the overall ability level of repeaters was lower than non-repeaters but the prior exposure caused the difficulty level of some anchor items in new the form was lower than the same items in the old form.

The response data were generated under four factors: sample size (7 levels) proportion of repeaters in the total sample (3 levels), the ability of repeaters (3 levels), the difference in form difficulty (2 levels). The dichotomous response data were generated by 3PL IRT model and polytomous responses were generated by graded response model (GRM: Samejima, 1972) to simulate the responses of a mixed-format test. Four equating approaches were performed and then compared with the criterion equating function - chained equipercentile equating.

3.2 Data Generation

The following subsections describe simulation procedure regarding examinee and the difference in test difficulty. The goal was to simulate ideal conditions (i.e., sufficient sample size, no difference in ability distribution and no drift in anchor) to the extremely poor conditions with very small sample size and relatively large drift.

3.2.1 Repeaters and Non-repeaters

According to the previous studies, repeaters tended to have lower overall performance than total group. In credentialing tests, most of the examinees retake the assessment because they fail the whole test or certain sub-tests at the previous administration. These repeaters might be different from the total group because their total score is lower than the cut-score (Andrulis, et al., 1978; Kim and Kolen, 2010; Puhan, 2011; Kim & Walker, 2012). In the real-world, the performance of repeaters is more

complicated than the performance in current simulation study. Some repeaters perform better than non-repeaters because they made progress after the first administration, while some of the repeaters have the same proficiency level as non-repeaters. The distribution of repeaters varies across tests, populations, sample size levels and so forth. However, the current study only considers the most common scenario in which repeaters have lower proficiency level than non-repeaters.

Based on the studies reviewed in the previous section, the range of sample size started from 10 to 3000. The current study chose the sample size of 20, 50, 100, 200, 300, 400 and 500 for the old test form; the criterion equating results were derived from a sample size of 5000. The large sample size of criterion equating function was used to ensure the accuracy of the equating results. The wide range of sample size would clearly display the change in equating results as the sample size increases. Although most of the studies have shown that repeaters have the lower mean score but similar variation, few studies compared the overall performance on a θ scale. In equating studies, groups differed 0.1 standard deviation units in ability can produce large bias in equating. The mean difference over 0.25 was considered extremely large and would yield drastic impact on equating accuracy (Wang, Lee, Brennan, & Kolen, 2008; Sunnassee, 2011). In the current study, non-repeaters who took the old form test were generated from a standardized normal distribution with a mean of 0 and variance of 1, denoted as $\theta_{NR} \sim N$ (0, 1). Within each sample size condition, examinees were categorized into two groups: repeaters and non-repeaters. Because the previous literature did not examine repeater's ability on θ scale, the current study included different levels of repeater ability and investigates which ability level might substantially influence the equating results. Three

repeater subgroups were generated: repeater group 1 (θ_{R1})– generated with a mean ability of -0.5 and standard deviation (SD) of 1, representing a slightly lower ability level; repeater group 2 –generated with a normal distribution of $\theta_{R2} \sim N$ (-1.0, 1), denoting a moderate lower ability; repeater group 3 –generated with a normal distribution of $\theta_{R3} \sim N$ (-1.5, 1), representing a substantially lower proficiency than total group. Under each sample size level, examinees in the new form consisted of certain proportion of repeaters from one repeater group ($\theta_{R1} \sim N$ (0, -0.5), $\theta_{R2} \sim N$ (-1.0, 1), or $\theta_{R3} \sim N$ (-1.5, 1) and non-repeaters. The next step is to combine repeaters from the old form with another group of new test takers to form the total sample of new test form. These non-repeaters were drawn from a standardized normal distribution of $\theta_{NR} \sim N$ (0, 1) with corresponding levels of sample size (20, 50, 100, 200, 300, 400 and 500). As a result, the numbers of examinee taking new form was larger than that of the examinees who take the old form, which is a common situation in practice. In Rogers and Radwan (2015), equating with 25%-35% of repeaters produced significantly increasing values of RMSE. In this study, 33% and 53% of examinees in old form were repeaters, resulting in 25% and 35% of repeaters in the new form test.

3.2.2 Test Difficulty

The length of the test was fixed to 36 multiple-choice (MC) items and 4 constructed-response (CR) items with a total score of 44. To equate the new form to the old form, 12 MC items were selected as anchor items. That was 24 unique (non-anchor) MC items and 4 unique (non-anchor) CR items. The items were obtained from a large-scale assessment. This simulation study included two pairs of test forms (Form 1 and Form 2). The summaries of item parameters are displayed in Table 3.1, Table 3.2 and

Table 3.3. The item parameters are displayed in Table A.1, Table A.2, and Table A.3. The test information functions of Form 1 and Form 2 are depicted in Figure 3.1 and Figure 3.2, respectively. The summary statistics of the raw score of two forms are presented in Appendix B. Form 1 test had no problematic anchor items while in Form 2, 6 anchor items in new form had lower difficulty than corresponding anchor items in the old form. However, these problematic anchor items were only easier to repeaters. The reason for using anchor items with lower difficulty level was to simulate a scenario that anchor items appear easier in new form only to repeaters who took the exact reference form. By decreasing the difficulty of anchor test, repeaters had better performance than first-time examinees only in anchor tests yet the total score on non-anchor items was still lower than the non-repeater group. The difference in anchor test difficulty was 0.50 standard deviation (SD) unit but the difference in overall difficulty was 0.17 SD unit. In the current study, the difference in test difficulty was mainly attributed to the difference in anchor tests, which was rarely considered in previous research. In most of the studies, it was more common to manipulate the test difficulty on unique items.

Table 3.1. Form 1 (no problematic anchor items)

	Old Form (*Y*)			New Form (*X*)			Anchor Form (*V*)		
	a_1	b_1	c_1	a_2	b_2	c_2	a	b	c
Mean	1.19	-0.15	0.19	1.21	-0.16	0.20	1.17	-0.17	0.16
SD	0.43	0.42	0.07	0.46	0.43	0.07	0.43	0.30	0.06
Min	0.53	-1.16	0.03	0.53	-1.16	0.03	0.53	-0.74	0.03
Max	2.03	0.71	0.33	2.07	0.71	0.33	2.03	0.41	0.28

Table 3.2. Form 2 (6 problematic anchor items)

	Old Form (*Y*)			New Form (*X*)			Anchor Form (*V*)		
	a_1	b_1	c_1	a_2	b_2	c_2	a	b	c
Mean	1.19	-0.15	0.19	1.21	-0.33	0.20	1.17	-0.67	0.16
SD	0.43	0.42	0.07	0.46	0.57	0.07	0.43	0.58	0.06
Min	0.53	-1.16	0.03	0.53	-1.72	0.03	0.53	-1.72	0.03
Max	2.03	0.71	0.33	2.07	0.71	0.33	2.03	0.03	0.28

Table 3.3. Item Parameters of CR items

	Old Form (*Y*)			New Form (*X*)		
	a_1	b_{11}	b_{12}	a_2	b_{21}	b_{22}
Mean	0.63	-1.24	1.24	0.63	-1.24	1.24
SD	0.19	0.09	0.09	0.19	0.09	0.09
Min	0.36	-1.34	1.11	0.36	-1.34	1.11
Max	0.88	-1.11	1.34	0.88	-1.11	1.34

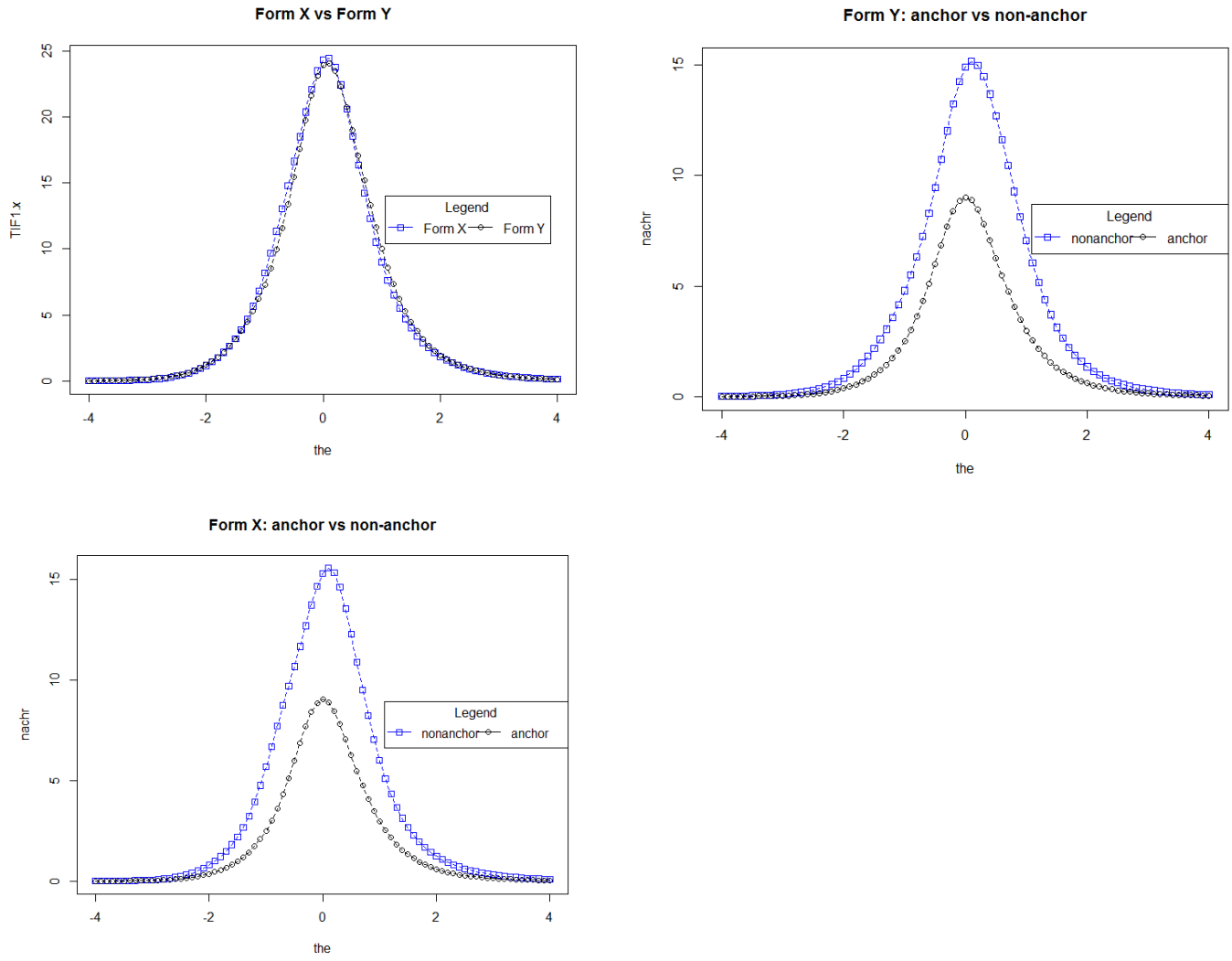


Figure 3.1. Test Information Function of MC items (Form 1)

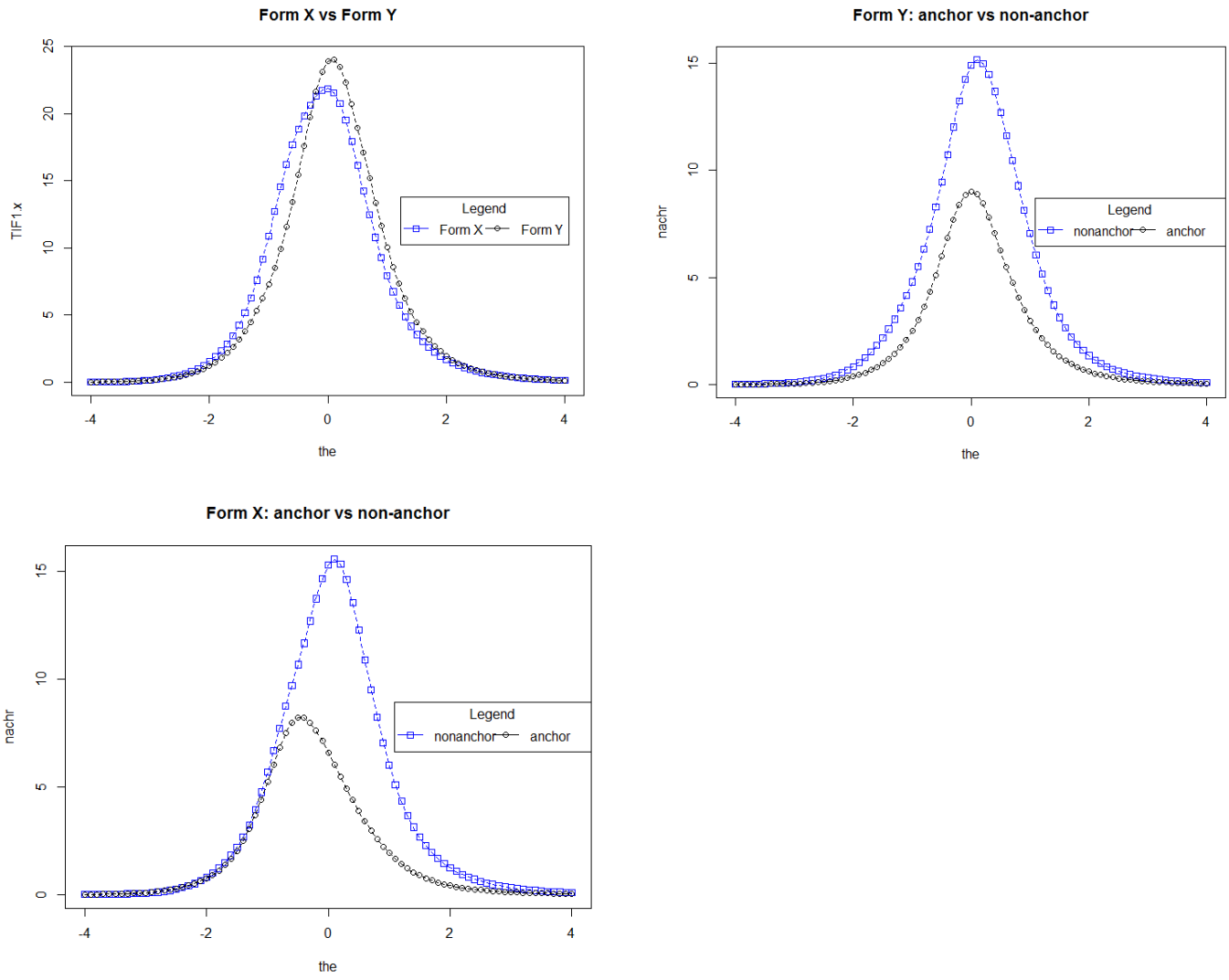


Figure 3.2. Test Information Function of MC items (Form 2)

3.2.3 Generation Model

The probability of giving correct response of binary data was generated by 3PL model presented in Equation (2.7), the correct responses were scored as "1" while the incorrect responses were scored as "0" Polytomous response data were generated by GRM, this model is expressed by equation (3.1)

$$P_{ijk}^{*}(\theta_j) = \frac{\exp[Da_i(\theta_j - b_{ik})]}{1 + \exp[Da_i(\theta_j - b_{ik})]} \quad (3.1)$$

where k=0, 1, 2,…, m_i . The k is the highest score a person can get on item i, and there are m+1 score categories. $P_{ijk}^{*}(\theta_j)$ denotes the conditional probability of an examinee j with ability level θ earning a score at or above k on item i. The a-parameter is a discrimination parameter that is constant across categories, b_{ik} is the threshold parameter for score k. The responses are scored depending on the k. Probability of each score category k on item i of examinee j can be given by

$$P_{ijk}(\theta_j) = P_{ijk}^{*}(\theta_j) - P_{ij(k+1)}^{*}(\theta_j), \quad (3.2)$$

In this study, the polytomous response data had three response categories. The incorrect response was scored as "0", the first correct response category was scored as "1"; the second correct response category was scored as "2". The probability of getting incorrect response is

$$P_{ij0}(\theta_j) = P_{ij0}^{*}(\theta_j) - P_{ij1}^{*}(\theta_j); \quad (3.3)$$

the probability for examinee i getting a score of "1" on item j is expressed as

$$P_{ij1}(\theta_j) = P_{ij1}^{*}(\theta_j) - P_{ij2}^{*}(\theta_j); \quad (3.4)$$

the chance of earning a score of "2" is

$$P_{ij2}(\theta_j) = P_{ij2}^{*}(\theta_j) - 0. \quad (3.5)$$

Data generation were performed using the computer program *R*, version 3.2.4 (Team, 2017), both binary response data and polytomous response data were generated using Monte Carlo simulation procedure over 100 replications (Harwell, Stone, Hsu, & Kirisci, 1996).

In sum, four factors were manipulated for data simulation: sample size (20, 50, 100, 200, 300, 400, 500 in old form), ability distribution of repeaters ($\theta_{R1} \sim N$ (-0.5, 1),

($\theta_{NR} \sim N$ (-1.0, 1) and $\theta_{R2} \sim N$ (-1.5, 1)), proportion of repeaters (25% and 35% in the new form) crossed with two levels of anchor test difficulty difference (0.50 and 0). At last, a special condition was simulated under each sample size where new form only has non-repeaters and there was no difference resulting from memorizing anchor items. As a result, there were 7*3*2*2+7= 91 simulation conditions.

3.3 Procedures

The procedure consists of four steps for IRT-based equating and three steps for classical equating. The first step was data generation based on the given item parameters in Appendix A using Monte Carlo simulation procedure. The first step was same across equating methods. Next, the data sets were prepared under three solutions conditions and one no-solution condition. In the first and third solution condition, all repeaters were removed; in the second solution, all examinees were included but problematic anchor items were excluded from the test for equating. In no solution conditions, no items or examinees were removed. For IRT equating (Rasch TSE), data were fitted to IRT models before performing IRT equating between old form and new form. The criterion equating function was derived from equipercentile equating with 5,000 examinees. The equipercentile equating function was frequently used as criterion equating function in NEAT design in previous research (e.g., Skaggs, 1995; Livingston & Kim, 2008; Kim & Livingston, 2010). Kolen and Brennen (2014) found a sample size of 1,500 is sufficient to perform equipercentile under NEAT design, this study used 5,000 examinees to prevent the irregularity of observed score distribution. Three classical equating approaches were performed using *R* package "*equate*" (Albano, 2016), two of them were

recently developed for small-sample equating, one was a conventional approach frequently used in previous studies.

For classical equating, the raw scores were converted without calibration. To perform IRT equating, the *R* package "*ltm*" (Rizopoulos, 2006) was used to fit Rasch model and PCM for the dichotomous and polytomous response, respectively. After equating process, the results were evaluated by evaluation criteria.

3.3.1 Parameter Calibration

Rasch model is a special case of 3PL model (equation 2.6) where all items have the same discrimination level (a = 1) and the zero guessing parameter (c = 0). The partial credit model (PCM: Masters, 1982) is an extension of Rasch model for polytomous response data. The PCM is presented as

$$p_{ijk}(\theta_i) = \frac{\exp[\sum_{h=1}^{k} D(\theta_i - b_j + d_{jh})]}{\sum_{g=1}^{m_j} \exp[\sum_{h=1}^{k} D(\theta_i - b_j + d_{jh})]}. \quad (3.6)$$

In the equation (3.6), $p_{ijk}(\theta_i)$ is the probability of responding in category k (k=0, 1,…, m) of item j, b_j is the item difficulty (location) parameter, and d_{j1}, d_{j2},…, d_{jh} are the category boundary (threshold) parameters for the item j. The d_{jh} defines how far the threshold is located from an item location b_j.The calibration procedures were conducted across all conditions using Rasch model and PCM before Rasch TSE.

In the current study, Rasch model was chosen over 3PL model because it can provide invariant item parameter and ability parameter as multiple-parameter IRT model but requires smaller sample size. One limitation of fitting Rasch model with data generated by multiple-parameter IRT model is misfit issues between model and data. Misfit may produce large parameter estimation error and thereby yielding biased and

inaccurate equating results. Applying Rasch TSE may retain the property of invariance but the property only holds if the model and data fit. The study explored if the benefits of Rasch TSE can offset the limitation of the misfit.

3.3.2 Nominal Weight Mean Equating

In NEAT design, a pair of test forms was administered to different groups of examinees. In most of the equating studies, the new form and the old form is represented by Form X and Form Y, and the common items between two forms is denoted as V. Suppose examinees take Form X are from Population 1 and examinees take Form Y are from Population 2, the equating function is derived from a single population involved both Population 1 and Population 2. This single population is referred to as the synthetic population (Braun & Holland, 1982). In nominal weight mean equating, the synthetic mean of Form X and Form Y are

$$\mu_S(X) = \mu_1(X) - w_2 \frac{K(X)}{K(V)}[\mu_1(V) - \mu_2(V)] \tag{3.7}$$

$$\mu_S(Y) = \mu_1(Y) + w_1 \frac{K(Y)}{K(V)}[\mu_1(V) - \mu_2(V)] \tag{3.8}$$

where μ refers to the mean, K indicates the number of items.

The weights w_1 and w_2 for Population 1 and Population 2 are computed based on number of examinees *N*, where

$$w_1 = \frac{N_1}{N_1 + N_2} \tag{3.9}$$

And

$$w_2 = \frac{N_2}{N_1 + N_2} \tag{3.10}$$

The formula for mean equating function is presented in equation (2.2)

$$y = m_Y(x) = x - [\mu_S(X) - \mu_S(Y)], \quad (2.2)$$

After substituting Equation (3.7) – (3.10) into Equation (2.2), the final nominal weight mean equating model is

$$m_Y(x) = x - \mu_1(X) + \mu_1(Y) + \frac{N_2K(X) - N_1K(Y)}{(N_2 + N_1) * K(V)} * [\mu_1(V) - \mu_2(V)]. \quad (3.11)$$

In equation (3.11), the estimate of equating function only includes the number of examinees of each group, number of items for the total tests and anchor test, mean of Form X from Population 1, Form Y from Population 2 and mean of anchor test from each population.

3.3.3 Circle-Arc Equating

The estimated equating function between alternate forms of circle-arc equating is an "arc curve" of a circle that is determined by passing through three prespecified score points from a new Form X and an old Form Y. These points are lower end-point, higher-point and a middle-point. The lower end-point (x_1, y_1) of the circle curve is determined by the lowest meaningful score on base Form Y (y_1) and new Form X (x_1), the upper end-point (x_3, y_3) is determined by the maximum possible score on base Form Y (y_3) and new Form X (x_3). If the equating design is a single group design or a random group design, the middle point (x_2, y_2) is determined by equating the mean score on the new form to the mean score on old form directly. If the equating design is non-equivalent group design performed on two groups of examinees, the middle point on old form $e_y(x)$ is treated as the equated mean score from Form X to Form Y. The middle point $e_y(x)$ can be

estimated by different approaches. Livingston and Kim (2009) estimated the middle point used chained linear equating, which is written as:

$$e_y(x) = l_y(x) = \mu_2(X) + \frac{\sigma_2(Y)}{\sigma_2(V)}\big(\mu_1(V) - \mu_2(V)\big) + \frac{\sigma_2(Y)}{\sigma_2(V)}\frac{\sigma_1(V)}{\sigma_1(X)}(x - \mu_1(X)). \quad (3.12)$$

In the equation, μ and σ indicate the means and standard deviation, examinees take the new Form X and the old Form Y and from Population 1 and Population 2, respectively. The chained linear equating does not consider population weights like nominal weights equating. For the circle-arc equating, the *x* in equation (3.12) is equal to $\mu_1(X)$, the simplified format that transforms the new form *x* to old *y* is

$$y_2 = \mu_2(X) + \frac{\sigma_2(Y)}{\sigma_2(V)}\big(\mu_1(V) - \mu_2(V)\big). \quad (3.13)$$

To obtain the middle points, 5 pieces of information are needed: the mean score of new form and old form anchor: $\mu_1(V)$ and $\mu_2(V)$, the standard deviation of old Form Y $\sigma_2(Y)$, the mean score of new form $\mu_2(X)$ and standard deviation of old anchor $\sigma_2(V)$.

There are two ways to use three points to determine an estimated equating curve. The first method is referred to as the symmetric circle-arc equating, the equating curve is the arc of the circle that constrained by two end-points and the middle point. When the new form is harder than old form and the middle point above a linear line connecting lower and upper point, the function of equating curve is denoted as

$$circ_Y(x) = y_c + \sqrt{r^2 - (x - x_c)^2}, \quad (3.14)$$

where the *r* is the radius of the circle curve, the point (x_c, y_c) is the center of the circle that passes three points. However, if the new form is easier than old form and the middle point is below the line, the function is formalized as

$$circ_Y(x) = y_c - \sqrt{r^2 - (x - x_c)^2}, \quad (3.15)$$

Because three points can constrain one circle, the estimation of radius and center of the circle is:

$$r^2 = (x_1 - x_c)^2 + (y_1 - y_c)^2 \tag{3.16}$$

$$x_c = \frac{(x_1^2 + y_1^2)(y_3 - y_2) + (x_2^2 + y_2^2)(y_1 - y_3) + (x_3^2 + y_3^2)(y_2 - y_1)}{2[x_1(y_3 - y_2) + x_2(y_1 - y_3)] + x_3(y_2 - y_1)]} \tag{3.17}$$

$$y_c = \frac{(x_1^2 + y_1^2)(x_3 - x_2) + (x_2^2 + y_2^2)(x_1 - x_3) + (x_3^2 + y_3^2)(x_2 - x_1)}{2[y_1(x_3 - x_2) + y_2(x_1 - x_3)] + y_3(x_2 - x_1)]} \tag{3.18}$$

Figure 3.3 shows the equating curve $circ_Y(x)$ passing through three points. The coordinates for three prespecified points are (5, 5), (12,14) and (20, 20).

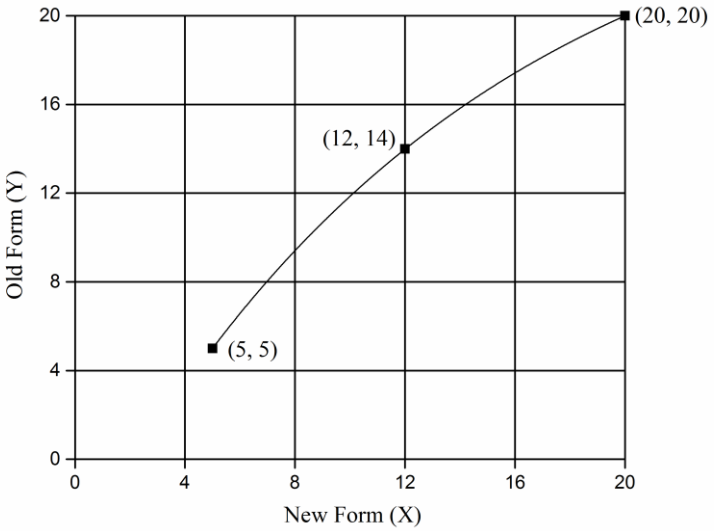


Figure 3.3. Systematic Circle-Arc Equating

The other method is simplified circle-arc equating, the equating curve is estimated by decomposing the function into a linear component *L(x)* that connects two endpoints (x_1, y_1) and (x_3, y_3) and a curvilinear component modeled by transformed points. The transformed points are obtained by subtracting the original prespecified points by the height of the linear component *L(x)*:

$$L(x) = y_1 + \frac{y_3 - y_1}{x_3 - x_1}(x - x_1) \tag{3.19}$$

$$y^* = y - L(x) \tag{3.20}$$

where y is the original end-point and y^* is the transformed point. In Figure 2, line $L(x)$ is found by constrained points (5,5) and (20, 20).

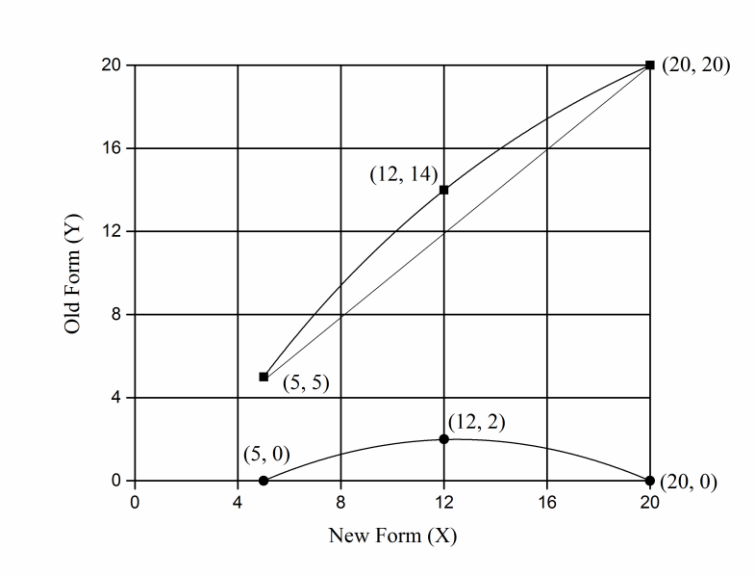


Figure 3.4. Simplified Circle-Arc Equating

The transformed points in Figure 3.4 are obtained by subtracting the $L(x)$ function. These transformed points are: transformed lower-point ($x_1 = 5$, $y^*_1 = 0$), the transformed middle point ($x_2 = 12$, $y^*_2 = 2$) and the transformed upper point ($x_3 = 20$, $y^*_3 = 0$). Next, the circle curve component is constrained by passing transformed points:

$$circ^*_Y(x) = y_c \pm \sqrt{r^2 - (x - x_c)^2}. \tag{3.21}$$

Finally, the circle curve function is the combination of the curvilinear component $circ^*_Y(x)$ and the linear function $L(x)$:

$$scirc_Y(x) = circ^*_Y(x) + L(x). \tag{3.22}$$

Symmetric and simplified equating methods yield similar equating results but the simplified method is computationally simpler and produces curves that are more similar

to the curves produced by equipercentile equating in large groups (Livingston & Kim, 2008; 2009). In the current study, simplified circle-arc equating were used.

3.3.4 Rasch Equating

After fitting Rasch model and PCM with generated data, estimated item parameters and estimated person parameters were obtained to build true score conversion table using statistical *R* package "*plink*" (Weeks, 2010). This procedure included two steps: rescaled item parameters and ability parameters from Form X to Form Y and then conducted TSE. The Stocking and Lord TCC method (1983) was used to determine the transformation constant *A* and *B*. Next, the scale transform was performed using the equation (2.8) – equation (2.11) for binary response data. Unlike dichotomous models where the items are linearly transformed by linking constants, the scale transformation for polytomous scored items requires the transformation on each category as well. In PCM, to perform the scale transformation from scale X to scale Y, the transformation of item *j* on category *k* is expressed as:

$$\delta_{Yjh} = A(b_{Xjh} - d_{Xjh}) = A(\delta_{Xjh}) + B \tag{3.23}$$

In the above equation, *A* and *B* are constants in the linear equation, δ_{Yjh} and δ_{Xjh} are the reparametrized difficulty of item *j* for category *h* on Scale Y and Scale X. The equation (2.9) and the equation (3.23) are analogous. After scale transformation, TSE process was performed to establish a conversion table where the Form Y true score is equivalent to a Form X true score for a given θ.

3.4 Evaluation Criteria

In the current study, six measures were computed to evaluate the bias and accuracy of equating: conditional equating bias, weighted average root mean squared bias (WRMSB), conditional standard error of equating (CSEE), weighted average standard error of equating (WSEE), conditional root mean squared error (RMSE), weighted average of RMSE (WRMSE). Conditional difference curves (CDC) was aimed to examine population invariance property. To evaluate the degree of population dependence, the difference that matters (DTM) was used to examine the degree of invariance at the cut-score point. The last evaluation criterion was decision accuracy (DA), which indicated how different factors impacted pass/fail decision at the individual level. Finally, all evaluation criteria would be reported in numerical and graphic formats. Statistical *R* package *"ggplot2"* (Wickham, 2016) was used to plot all the graphs.

3.4.1 Equating Bias and Accuracy

Equating bias is an index of systematic error of equating. The conditional bias at each score point is calculated by

$$Bias\ (x) = \frac{1}{M}\sum_{i=1}^{M}[\widehat{e_i}(x) - e(x)], \tag{3.24}$$

where M =100 is the total number of replication, x is a score point, $e(x)$ is the criterion equating function. The criteria equating function were estimated by chained equipercentile equating with 5000 examinees. The term $\widehat{e_i}(x)$ is the sample equating function that transforms scores of Form X to the score scale of Form Y in the *i*th Monte Carlo replication. The WRMSB was computed to indicate the overall measure of systematic errors across score range. The reason to use WRMSB was to prevent the

negative and positive values across score levels canceling each other. The formula for the measures is:

$$WRMSB = \sqrt{\sum_x r_x Bias^2(x)}. \quad (3.25)$$

The r_x is the ratio of sample size at certain score point *x* over total sample size.

SEE indicates the random error in equating that is due to sampling variability. The CSEE at score point *x* is computed as

$$CSEE\ (x) = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left[\hat{e}_i(x) - \bar{\hat{e}}(x)\right]^2}, \quad (3.26)$$

in which $\bar{\hat{e}}(x)$ is the average of the sample equating function over $M = 100$ replications.

Similarly, the WSEE across score ranges is defined as

$$WSEE = \sqrt{\sum_x r_x CSEE^2(x)}. \quad (3.27)$$

The RMSE and WRMSE are calculated as

$$RMSE\ (x) = \sqrt{\frac{1}{M}\sum_{i=1}^{M}[\hat{e}_i(x) - e(x)]^2}, \quad (3.28)$$

$$WRMSE(x) = \sqrt{\sum_x r(x) RMSE^2(x)}. \quad (3.29)$$

RMSE and WRMSE denote the combination of systematic and random errors.

3.4.2 Criteria for Equating Invariance

Conditional difference curves (CDC) quantify the degree of population variance. In NEAT data collection design, CDC is defined in observed score scale as

$$CDC(x) = e_{Pj}(x) - e_P(x). \quad (3.30)$$

In equation (3.40), $e_{Pj}(x)$ represents the equating function derived from group P_j, which is non-repeater groups in this study; while $e_P(x)$ is the equating function in the total sample at score point *x*.

The DTM was used to determine the acceptable level of violation to invariance property. The DTM represents a half unit of reported score unit, the scoring is based on a number of correct items which is equal to 0.5 in this study. The current study mainly concerned about the violation to invariance property at the cut-score point, which was the point determined the pass/fail decision. Values of CDC at a cut-score point below 0.5 range were considered as acceptable violation; otherwise, the violation was non-negligible.

3.4.3 Decision Accuracy

In addition to the indices of equating errors and bias, the DA was computed to indicate the decisions made based on individual rescaled observed/true score across different test conditions and equating approaches. To simulate the scenarios of credentialing tests, the current study only included two performance categories (pass/fail). The cut-score was 26 points, a raw score less than total score 26 would be considered as fail, a raw score equaled to or greater than 26 would be treated as pass. This cut score was selected because more examinees obtained a score in the middle of score scale than two ends of the score scale. If the cut point was located at two ends of the score range, the accuracy would be very high due to chance. To compute the "true classification", the cut-score was mapped from the observed score metric (x=26) to the θ scale through the TCC. Then each examinee was classified into "true" performance categories based on the cut-score on true score scale. The observed classification was made based on the equated score of simulation response data. Finally, DA was computed based on the classification of examinee's true ability and classification based on observed classification over 100 replications.

DA describes the degree to which actual classification agree with true classification. Table 3.4 gives an example of DA of an administration of the test. In Table 3.4, P_{PP} = 350/1000 = 35% defines the proportion of examinees categorized as passing between observed and true score equating; while P_{FF} defines the proportion of the examinees failed the test P_{FF} = 500/1000 = 50% between true and observed decision. The DA is computed as the sum of P_{PP} and P_{FF} , which results in $P_A = P_{PP} + P_{FF}$ = 85%.

Table 3.4. Number of Pass or Fail examinees

	Observed		
True	Pass	Fail	Total
Pass	P_{PP} = 350	P_{PF} = 50	$P_{P.}$ = 400
Fail	P_{FP} = 100	P_{FF} = 500	$P_{F.}$ = 600
Total	$P_{.P}$ = 450	$P_{.F}$ = 550	$P_{..}$ = 1000

3.5 Summary

In the current study, response data were simulated by manipulating two types of factors. One was related to examinee characteristics: sample size, ability distribution, the proportion of repeaters. The ability of non-repeaters followed a standardized normal distribution. There were three levels of ability distribution for repeaters, it was assumed that all of them reviewed the same set of anchor items in the reference form. By doing so, they may have better performance on anchor tests but lower performance on unique items. The other type of factor, which was relevant to the test characteristics of new form, was manipulated by the number of problematic anchor items. The items that were memorized by repeaters appeared easier only to repeaters; therefore, the difficulty level of anchor test, as well as total test, was lower than the true item difficulty for repeater groups. After data generation, each pair of data set were equated using one traditional small-sample equating technique, two recently developed technique, and the IRT TSE

approach. Because each method has limitations and strengths, it is hard to find an equating method solves all the problems. Nevertheless, the study would show which equating techniques can provide more desirable equating results under certain conditions. Table 3.5 summarizes all conditions of the current study. Under the following conditions, it was very likely that the large anchor test difficulty differences and difference between repeaters and non-repeaters produce high biased equating results; however, the goal was to simulate a "worst-case scenario" and examine how different equating performed under such conditions.

Table 3.5. Conditions in the Simulation Study

Sample Sizes	N_{old}= 20, 50, 100, 200, 300, 400, 500
Ability	$\theta_{NR} \sim N$ (0, 1) for non-repeaters, $\theta_{R1} \sim N$ (-0.5, 1), $\theta_{R2} \sim N$ (-1.0, 1), and $\theta_{R3} \sim N$ (-1.5, 1)
Difference in Anchor Test Difficulty	$b_1(V) - b_2(V) = 0$, $b_1(V) - b_2(V) = 0.50$
Proportion of Repeaters	0%,25% and 35% for new form
Equating Methods	identity, nominal weight mean, circle-arc, Rasch and PCM TSE

Three solutions were proposed to mitigate repeater effects. The first solution was to remove all repeaters before equating and then apply the equating function derived from the non-repeaters sample to total sample group. This solution was commonly used in practice but was criticized for two reasons. Firstly, excluding repeaters would reduce sample size and result im large standard errors. In addition, people may argue that equating invariance property does not hold because the equating relationship is obtained from a subgroup but used for the total group. Therefore, the first solution was evaluated by CDC, which focused on the difference between non-repeaters and total sample, as well as equating accuracy and DA. The second solution was eliminating the anchor items

that were memorized by examinees. Because repeater retained in the total sample, the evaluation criteria only involved indices evaluating bias, standard errors, the overall equating accuracy and DA. The last solution was to use IRT TSE which can also be nested within the first and second solution, which were removing all repeaters and applied IRT TSE, discarding drifted items but retaining all repeaters using IRT TSE. The final step was to perform equating without any solutions and compare whether these solutions can improve equating dramatically. Table 3.6 summarizes all solutions and corresponding equating methods and evaluation criteria. It should be noticed that the solution 3 can also be embedded within solution 1 and solution 2 when the equating method was Rasch TSE.

In this study, three data management approaches were prepared before equating, the last row in Table 3.6 represents no solution condition. All equating procedures were performed on each data set. Certain evaluation criteria corresponding to each solution were listed in Table 3.6.

Table 3.6. Equating Methods and Criteria for Each Solution

Solutions	Equating	Evaluation Criterion	Data Management
Solution 1: Removing repeaters	Identity, Nominal Weight Mean, Circle-Arc, Rasch and PCM TSE	CDC, bias, standard error of equating, RMSE, DA	Removing repeaters before equating
Solution 2: Discarding problematic anchor Items	Identity, Nominal Weight Mean, Circle-Arc, Rasch and PCM TSE	bias, standard error of equating, RMSE, DA	Excluding problematic anchor before equating
Solution 3: IRT equating	Rasch and PCM TSE	bias, standard error of equating, RMSE, DA	Removing repeaters, excluding problematic anchors or retaining all responses
No solution	Identity, Nominal Weight Mean, Circle-Arc	bias, standard error of equating, RMSE, DA	Retaining all responses

CHAPTER 4

RESULTS

Recall that the factors manipulated in the current research were sample size, repeater ability level, anchor test difficulty (item difficulty drift), proportion of repeaters, repeater effect solutions, and equating methods. Three overarching questions were asked:

1. Under the same test conditions and using same small sample equating techniques, how do different repeater effects solutions impact the equating results?
2. How do different small sample equating techniques impact the equating results?
3. What are the practical implications of this study?

The first research question mainly investigates solutions to mitigating repeater effects, the second research question emphasizes the comparison among small sample equating techniques while the last research question asks the practical implications of equating results and recommendations regarding equating design under small sample conditions. The first and second research question would be answered concurrently according to the conditional as well as overall equating bias, equating errors, population invariance measurement. The third question is answered based on the results of decision accuracy (DA) The recommendations to small volume testing programs are fully discussed in the next chapter.

The results chapter has 8 sections corresponding to each evaluation criterion: conditional bias, weighted average root mean bias (WRMSB), conditional standard error of equating (CSEE), weighted average standard error of equating (WSEE), conditional root mean average squared error (RMSE), weighted average of RMSE (WRMSE), conditional difference curve (CDC) and decision accuracy (DA). The conditional

equating results are reported because they show the patterns of equating bias and error across score scale. For the conditional difference curves, the magnitude of CDC was evaluated by the difference that matters (DTM) at the cut-score point (x = 26). Under each section, the first subsection analyzes the results when equating with non-problematic (non-drifted) anchor condition; the second analyzes equating with a problematic (drifted) anchor; the final subsection discusses the difference between repeater means. In terms of the way to display equating results, all results are presented in tables and figures. The tables in Appendix B show summary statistics while the graphical approach mainly depicts the patterns of equating bias and errors across score scales or test conditions.

4.1 Effects on Conditional Equating Bias

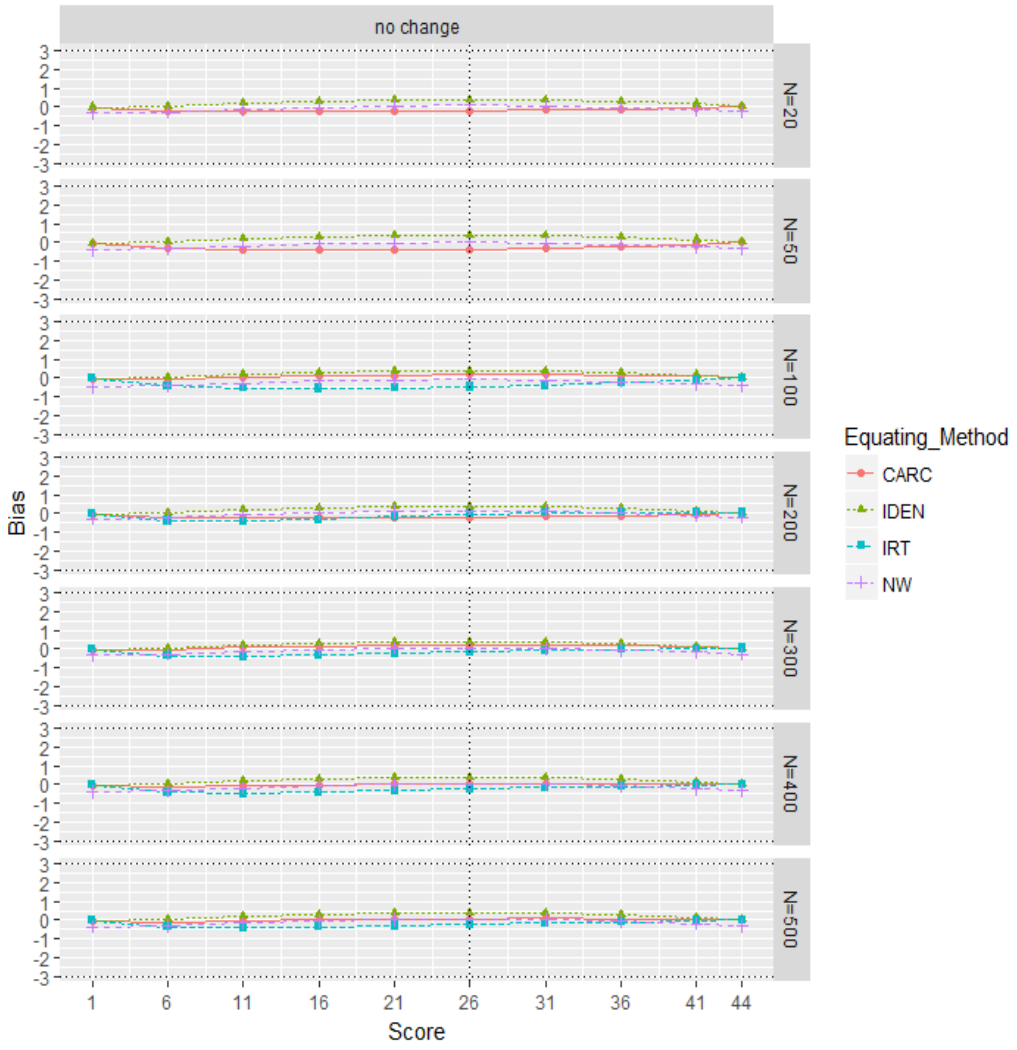
This section presents the conditional equating bias. The patterns of equating bias are depicted in Figure 4.1 to Figure 4.7. To highlight the comparison among small sample equating approaches, the distribution of repeaters was fixed to $\theta_{R3} \sim N$ (-1.5, 1) from Figure 4.1 to Figure 4.6. In Figure 4.7, 21 panels are displayed to investigate if different repeater means led to different equating bias conditional using one equating technique. The circle-arc equating was selected because this observed score equating approach can be applied to all sample size levels. Figure 4.1 to Figure 4.3 display the equating bias when there was no drift in anchor while Figure 4.4 to Figure 4.6 show the equating bias resulted from the problematic anchor. In each figure, the panels in the same row refer to the same sample size level and the panels in the same column represent the same repeater effect solution, which also denotes the data management strategies were made before equating was performed. Figure 4.2 and Figure 4.3 presented in the first subsection only have two columns because the difficulty level of new form did not contain any drift, and

there was no need to exclude problematic anchor items. The first column represents the repeater effect solution where the repeater responses were removed from the original data. The second column displays the results where all responses were retained. The second subsection (4.1.2) presents the results under two solutions to mitigate repeater effects (removing repeaters and excluding problematic anchor items) and the results when no change was made to the data set. The proportion of repeater was fixed in each figure but differs between figures. Under each section, the impact of equating method, sample size, and repeater effect solution are described within and between repeater proportion conditions. The last subsection (4.1.3) analyzes if repeater distribution has an influence on conditional bias using circle-arc equating with 35% repeaters and problematic anchor items. This condition is displayed as an example of the "worst case scenario" where the magnitudes of the anchor drift and repeater effects were the largest. If the equating results were not substantially different across repeater means under this scenario, it might be safe to conclude that repeater mean has a small influence on equating bias.

4.1.1 Non-problematic Anchor

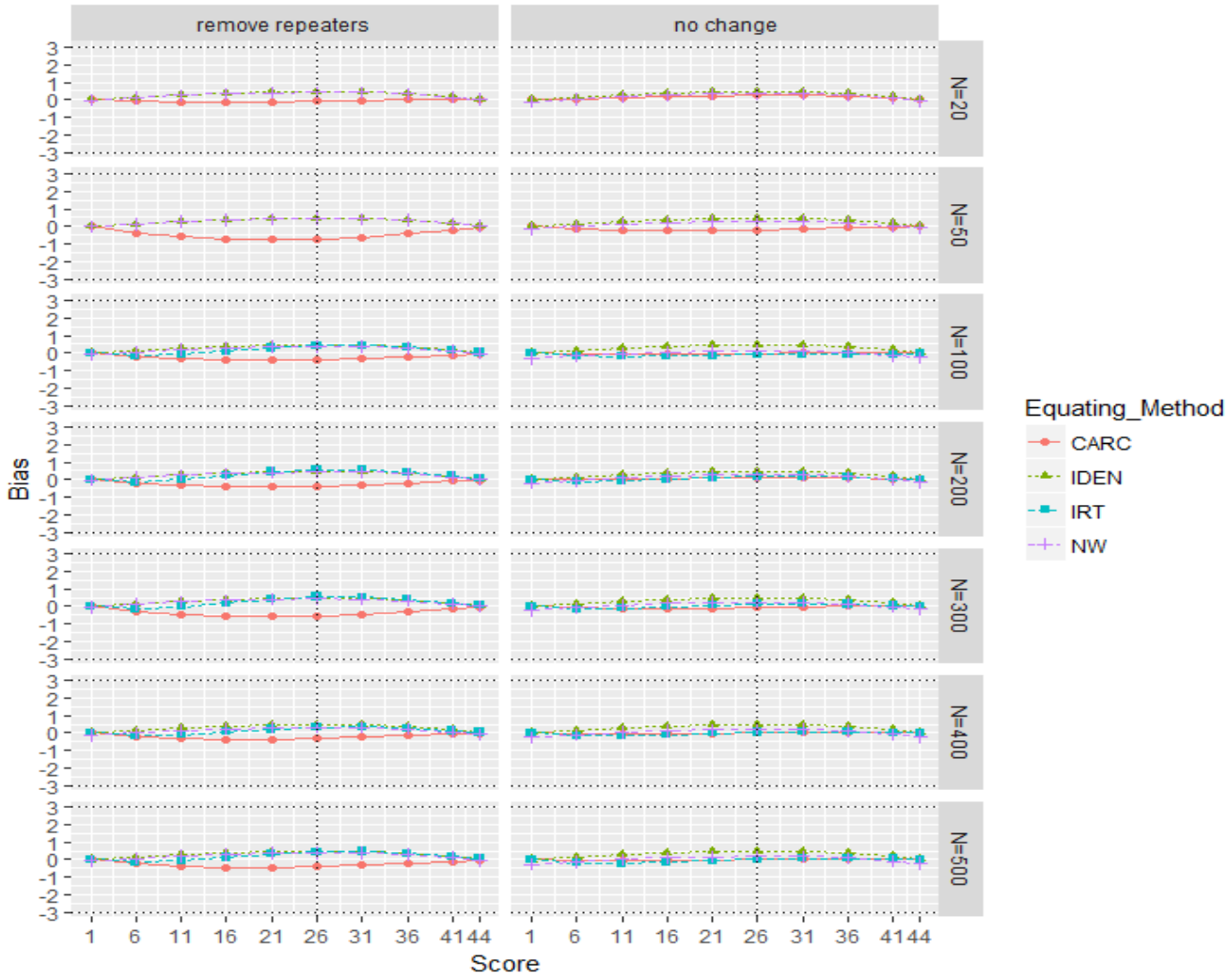
Several observations can be made from Figure 4.1 to Figure 4.3. Firstly, the conditional bias tended to be greater in the middle and smaller at the ends of the score scale. This trend was less noticeable under non-repeaters conditions. In Figure 4.1, which was the condition with no drift in the anchor and no repeaters, the overall patterns of conditional bias were not substantially different across all test conditions. However, when repeater proportion were 25% and 35% (Figure 4.2 and Figure 4.3, respectively), circle-arc equating was likely to produce negative bias while other equating techniques resulted

in positive bias values. The findings regarding equating methods reveal that the divergence between circle-arc and other equating methods was augmented as a proportion of repeaters increased from 0% to 35%. However, the gap between equating techniques was minimized if repeaters were not removed from original data set (see the last panel of the second column). With respect to the influence of sample size, the patterns of bias were not dramatically differed across size levels yet there was more variability among equating techniques at sample size level of 20.



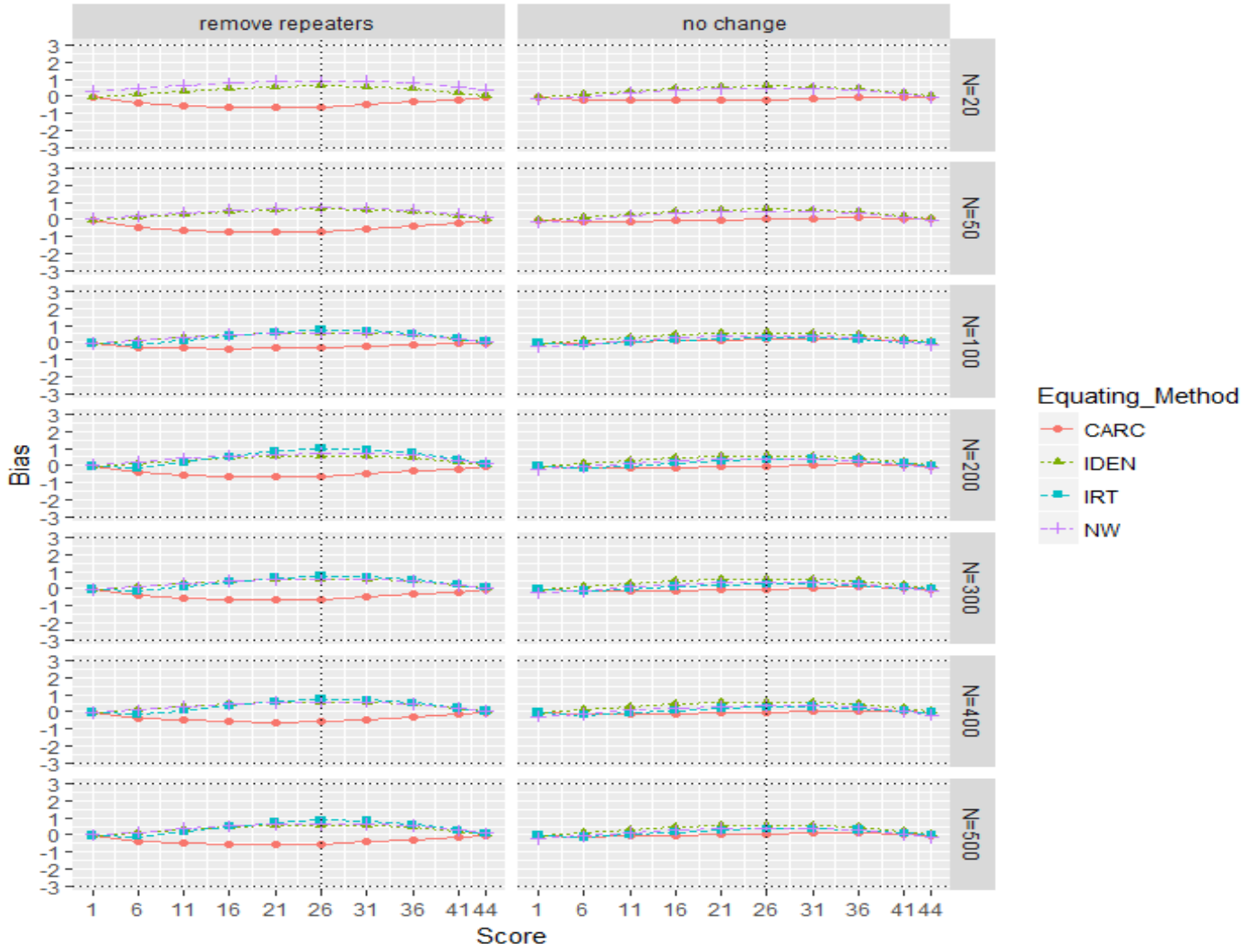
Note. Distribution of repeaters: $\theta_{R3} \sim N$ (-1.5, 1)

Figure 4.1. Bias of Non-problematic Anchor Test with 0% Repeaters by Equating Methods



Note. Distribution of repeaters: $\theta_{R3} \sim N$ (-1.5, 1)

Figure 4.2. Bias of Non-problematic Anchor Test with 25% Repeaters by Equating Methods

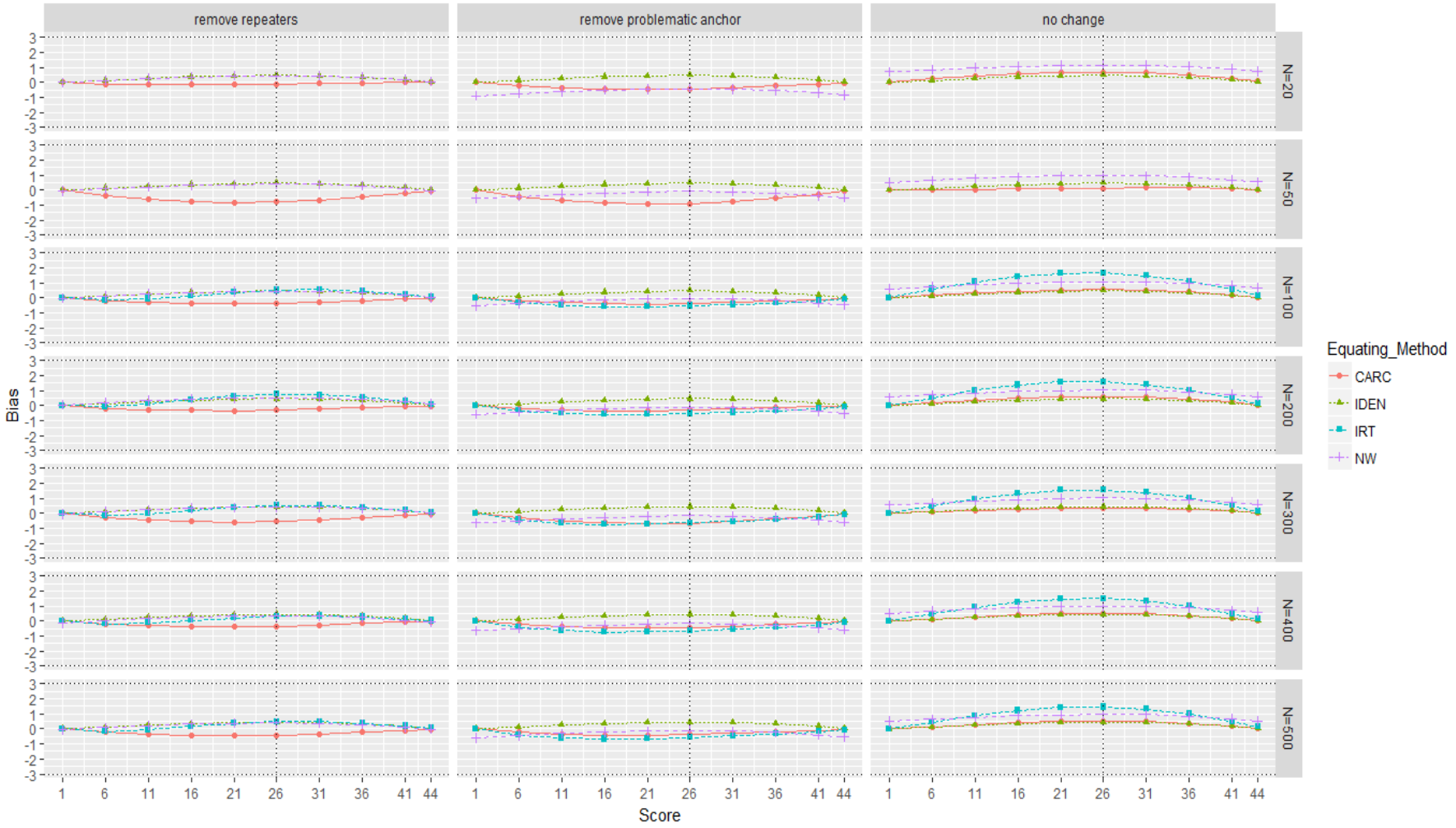


Note. Distribution of repeaters: $\theta_{R3} \sim N$ (-1.5, 1)

Figure 4.3. Bias of Non-problematic Anchor Test with 35% Repeaters by Equating Methods

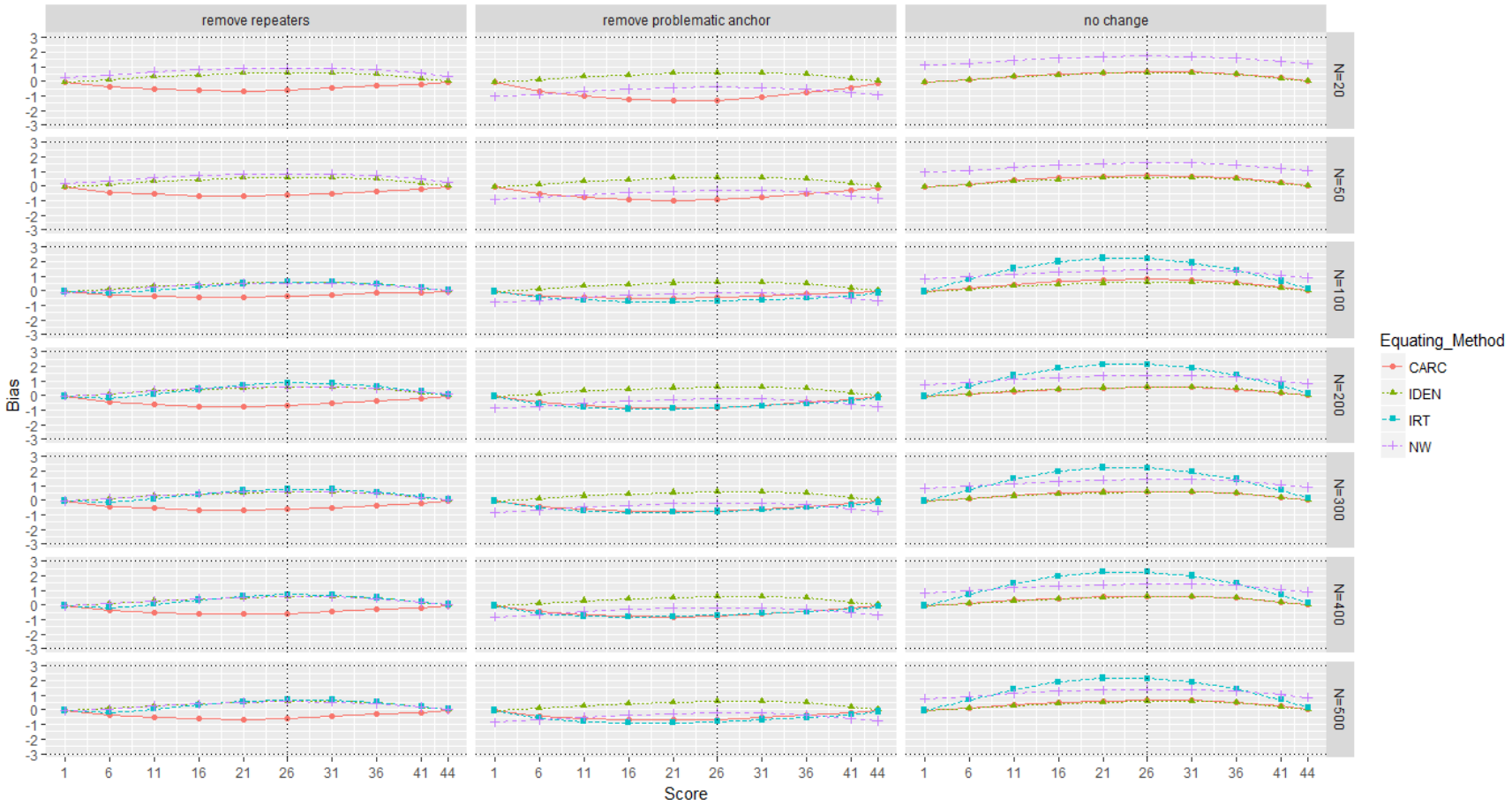
4.1.2 Problematic Anchor

Figure 4.4 to Figure 4.5 display the conditions when equating with a problematic anchor. The patterns of conditional bias have some similarities between Figure 4.1 - Figure 4.3 and Figure 4.4 – Figure 4.5. Firstly, within one repeater effect solution, bias produced by different equating techniques did not reduce from small to large sample size levels. The other similarity was related to the impact of repeater proportion on the conditional bias. When the proportion of repeaters increased from 0% to 35%, the disagreement among equating techniques was magnified. In terms of the dissimilarity between anchor test conditions, circle-arc equating was not the only equating method that resulted in negative conditional bias. Rasch equating also had a negative bias if the drifted anchor was excluded from equating. A closer look at Rasch equating results shows that Rasch equating yielded the highest positive bias if both problematic anchor and repeaters were retained (see last column of the panel). Under this condition, circle-arc and identity equating produced positive and least biased equating results. By holding the same proportion of repeaters, Rasch equating and nominal weight mean equating were more likely to produce a conditional bias that was substantially larger from circle-arc equating and identity equating if all repeater responses and anchor items were retained. This finding would be further confirmed by analyzing overall bias in following sections



Note. Distribution of repeaters: $\theta_{R3} \sim N$ (-1.5, 1)

Figure 4.4. Bias of Problematic Anchor Test with 25% Repeaters by Equating Methods

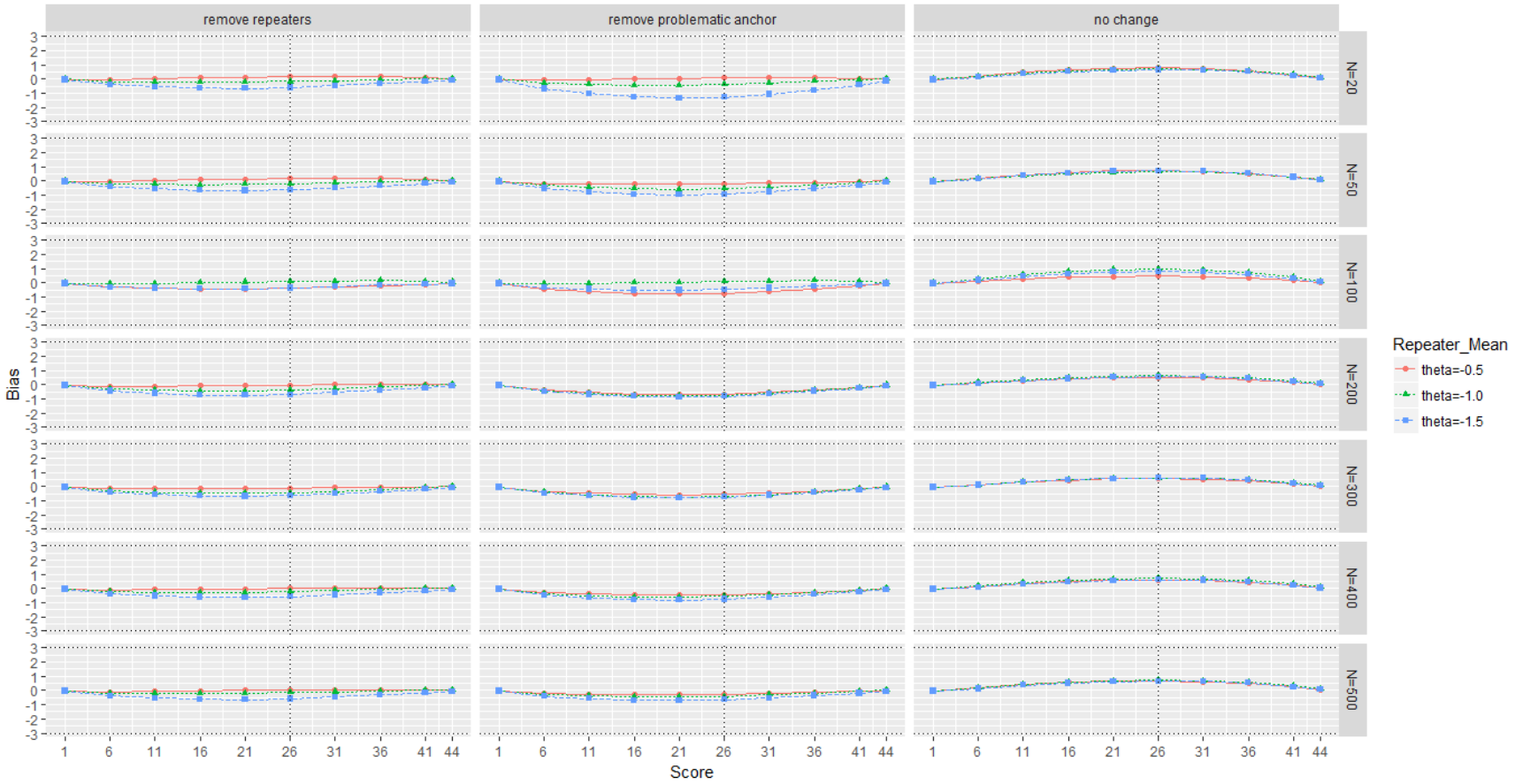


Note. Distribution of repeaters: $\theta_{R3} \sim N$ (-1.5, 1)

Figure 4.5. Bias of Problematic Anchor Test with 35% Repeaters by Equating Methods

4.1.3 Repeater Mean

Regarding the finding related to repeater mean, Figure 4.6 compares how conditional bias differs across three repeater distributions. The larger variability between repeater distributions was found at smaller sample size levels. The disagreement between repeater means was most noticeable under removing problematic anchor at sample size levels of 20 and 50. When $N \leq 50$, equating results were less biased if the repeater mean was close to zero. When $N > 50$, the bias resulted from different repeater distributions were very consistent across sample size levels and solutions. Compared with Figure 4.1 to Figure 4.5, repeater mean had a relatively smaller influence than equating techniques, even under the condition with largest repeater effects and item difficulty drift.



Note. Equating Method: Circle-Arc Equating

Figure 4.6. Bias of Problematic Anchor Test with 35% Repeaters by Repeater Mean

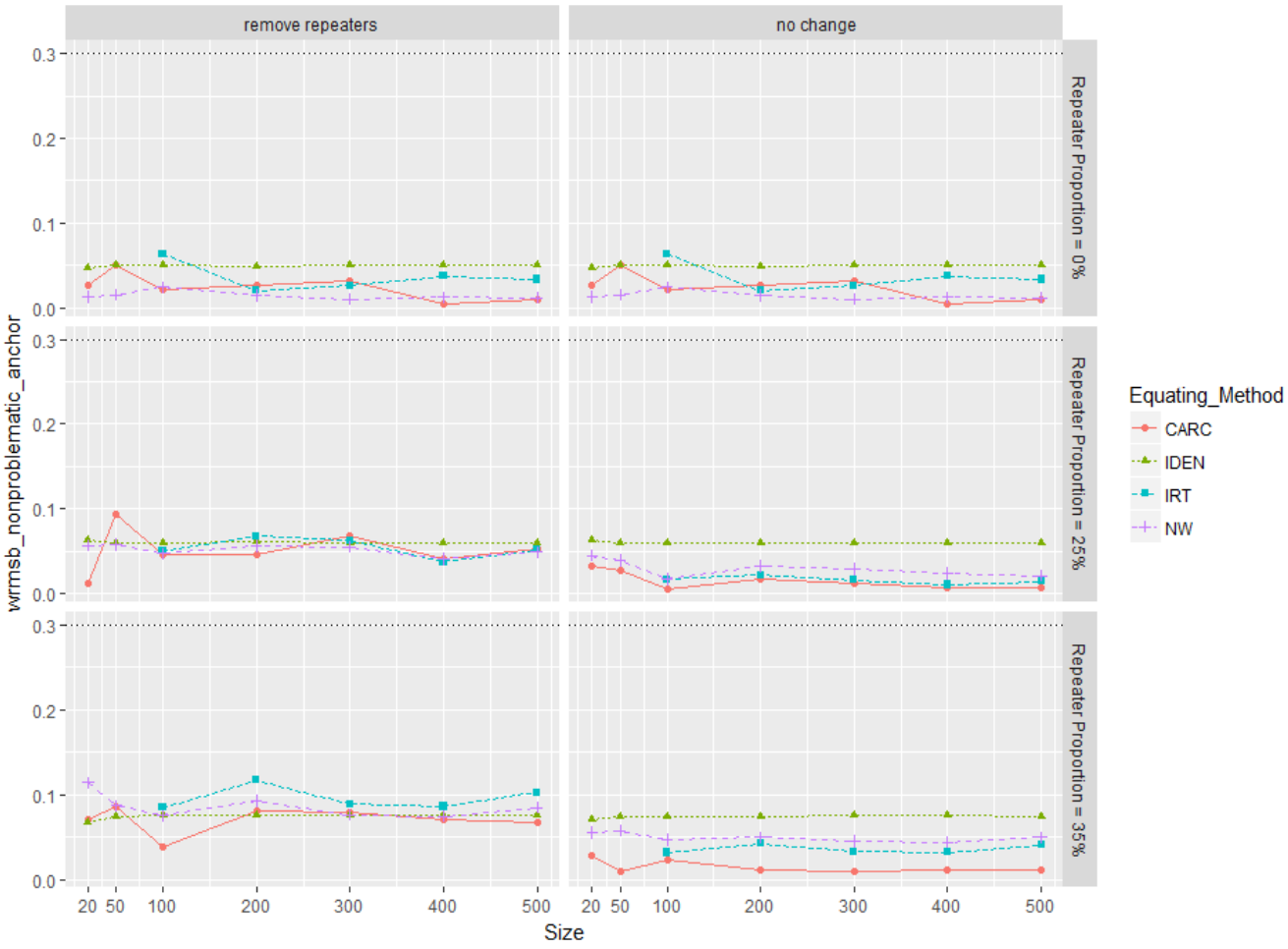
4.2 Effect on WRMSB

WRMSB indicates the overall equating bias by considering the proportion of examinees at each score point. The current section consists of two subsections. The first subsection (4.2.1) focuses on analyzing overall equating bias under the equating condition with a non-problematic anchor. The second subsection examines if WRMSB changed dramatically under the equating condition with a problematic anchor. The patterns of WRMSB are presented in Figure 4.7 to Figure 4.9. Figure 4.7 and Figure 4.8 emphasize how WRMSB differs between equating anchors while Figure 4.9 is an example showing if the WRMSB values are different between three repeater distributions. In each figure, line charts in the same column represent the WRMSB under same repeater effect solutions while the charts in the same row represent the same proportion of repeaters. The summary statistics are listed at the Table B6 and Table B7 in Appendix B.

4.2.1 Non-problematic Anchor

Figure 4.7 has 6 line charts. The line charts in the first column show the WRMSB under removing repeater condition and the charts at the second column show the WRMSB under retaining repeaters condition. This figure only has six charts because the equating was performed with non-problematic anchor and there is no need to display removing problematic anchor solution. According to the figure, the magnitude of WRMSB was generally low across all conditions. The mean of WRMSB across sample size levels ranged from 0.03 (SD = 0.02) at sample size of N = 400 to 0.05 (SD = 0.03) with the sample size of 20. In the figure, the WRMSB at small sample size level was

higher and has more variability. In the two graphs, retaining all repeaters results in slightly higher bias then excluding repeaters. The results might be intuitive. However, the true equating function was derived from a sample consists of repeaters and non-repeaters. This leads solution of retaining repeater solution produced an estimated equating function closer to true equating function. The identity equating method was likely to produce a slightly higher WRMSB if repeater responses retained. The mean WRMSB produced by circle-arc equating, identity equating, Rasch equating and nominal weight mean equating was 0.03 (*SD* = 0.03), 0.06 (*SD* = 0.01), 0.03 (*SD* = 0.02) and 0.03 (*SD* = 0.02), respectively. This may confirm that identity equating produced slight higher bias but the differences between equating methods were trivial. The means of WRMSB under "removing repeaters" and "retaining repeaters" conditions were 0.04 (*SD* = 0.03) and 0.03 (*SD* = 0.02), respectively. Visual inspection to figures and summary statistics reveals that the proportion of repeaters did not play an important in influencing the WRMSB. The mean and standard deviation for 0%, 25% and 35% condition were 0.03 (*SD* = 0.02), 0.04 (*SD* = 0.02), 0.04 (*SD* = 0.03), respectively.



Note. Distribution of repeaters: $\theta_{R3} \sim N$ (-1.5, 1)

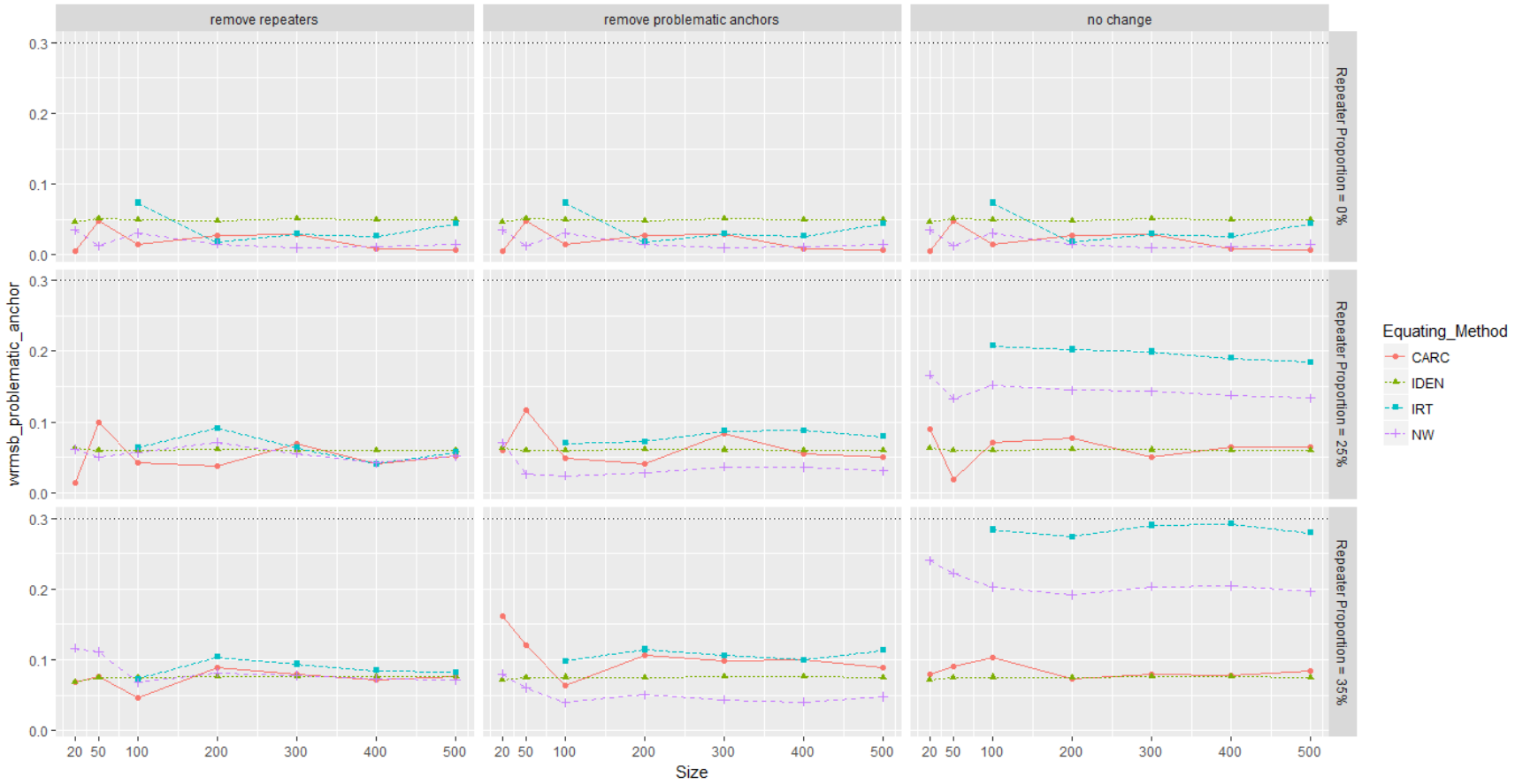
Figure 4.7. WRMSB of Non-problematic Anchor Test by Equating Methods

4.2.2 Problematic Anchor

Figure 4.8 presents WRMSB when equating with drifted anchor items. The charts at the first column show the WRMSB under "removing repeaters" solution, the second column displays the charts of "removing problematic anchor" solution and the charts at the third column show the WRMSB under "retaining repeaters and problematic anchor" condition. The magnitude of WRMSB was generally low (smaller than 0.1) under the "removing repeater condition" with a mean of 0.04 (*SD* = 0.03). The magnitude of the overall bias was higher under the other solutions with a mean of 0.06 (*SD* = 0.03) and 0.09 (*SD* = 0.07) for "excluding problematic anchor" and "retaining all items and repeaters" condition, respectively. The larger amount of bias may be resulted from the interaction of repeater proportion, equating techniques and repeater effect solutions. The magnitude of WRMSB increased as the proportion of repeater increased. The influence of repeater proportion was more significant if no repeater responses or drifted anchors were removed (see the charts in the third column). Also, this condition shows that Rasch equating and nominal weight mean equating can produce distinctly higher bias than circle-arc and identity equating. For example, the value of WRMSB produced by Rasch equating was approximate to 0.3 whereas circle-arc and identity equating yielded bias below 0.1 when 35% repeaters were all retained before equating procedure. Overall, the mean WRMSB produced by circle-arc equating, identity equating, Rasch equating and nominal weight mean equating is 0.05 (*SD* = 0.04), 0.06 (*SD* = 0.01), 0.10 (*SD* = 0.08) and 0.06 (*SD* = 0.06). The mean difference in WRMSB between equating methods can be strongly impacted by the data management strategies. The mean WRMSB ranges between 0.05 (*SD* = 0.04) at sample size 50 to 0.07 (*SD* = 0.07) at sample size of 20. As a

result, there was no evidence indicating WRMSB significantly reduced with larger sample size.

The overall means for non-problematic and problematic anchor equating condition were 0.04 (*SD* = 0.02) and 0.06 (*SD* = 0.05), respectively. In sum, the problematic anchor test caused higher overall bias and larger variation. The higher bias in Figure 4.8 was essentially caused by the equating conditions presented at the last two charts under "retaining all response data" condition. This may also indicate Rasch equating and nominal weight mean equating were more sensitive to drift in anchor test and a large proportion of repeaters. These two equating techniques were likely to result in highly biased results if no actions were taken to mitigate repeater effects or exposed anchors.

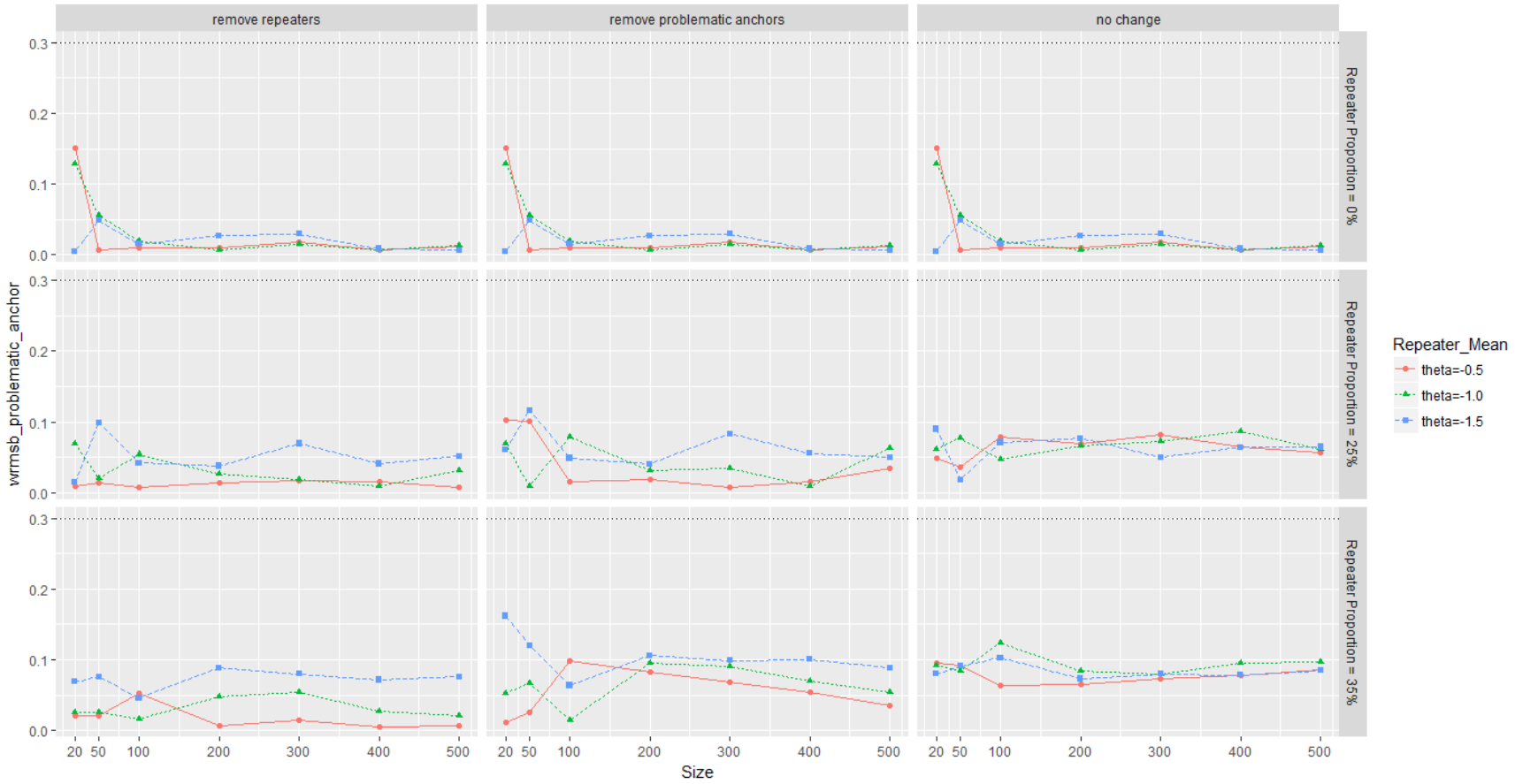


Note. Distribution of repeaters: $\theta_{R3} \sim N$ (-1.5, 1)

Figure 4.8. WRMSB of Problematic Anchor Test by Equating Methods

4.2.3 Repeater Mean

For non-problematic anchor condition, the mean and standard deviation for $\theta_{R1} \sim N$ (-0.5, 1), $\theta_{R2} \sim N$ (-1.0, 1) and $\theta_{R3} \sim N$ (-1.5, 1) were 0.03 (SD = 0.02), 0.04 (SD = 0.02), 0.04 (SD = 0.02), respectively. Under problematic anchor condition, the mean and standard deviation for $\theta_{R1} \sim N$ (-0.5, 1), $\theta_{R2} \sim N$ (-1.0, 1) and $\theta_{R3} \sim N$ (-1.5, 1) were 0.05 (SD = 0.05), 0.06 (SD = 0.051), 0.07 (SD = 0.06), respectively. In Figure 4.9, lines representing different repeater mean were consistent across sample size levels. Although sample size level N = 20 has more variation among three mean distributions, other test conditions still implied high agreement between means. The summary statistics may further confirm that drifted anchor has a stronger influence on enlarging bias than repeater mean and sam ple size.



Note. Equating Method: Circle-Arc Equating

Figure 4.9. WRMSB of Problematic Anchor Test by Repeater Mean

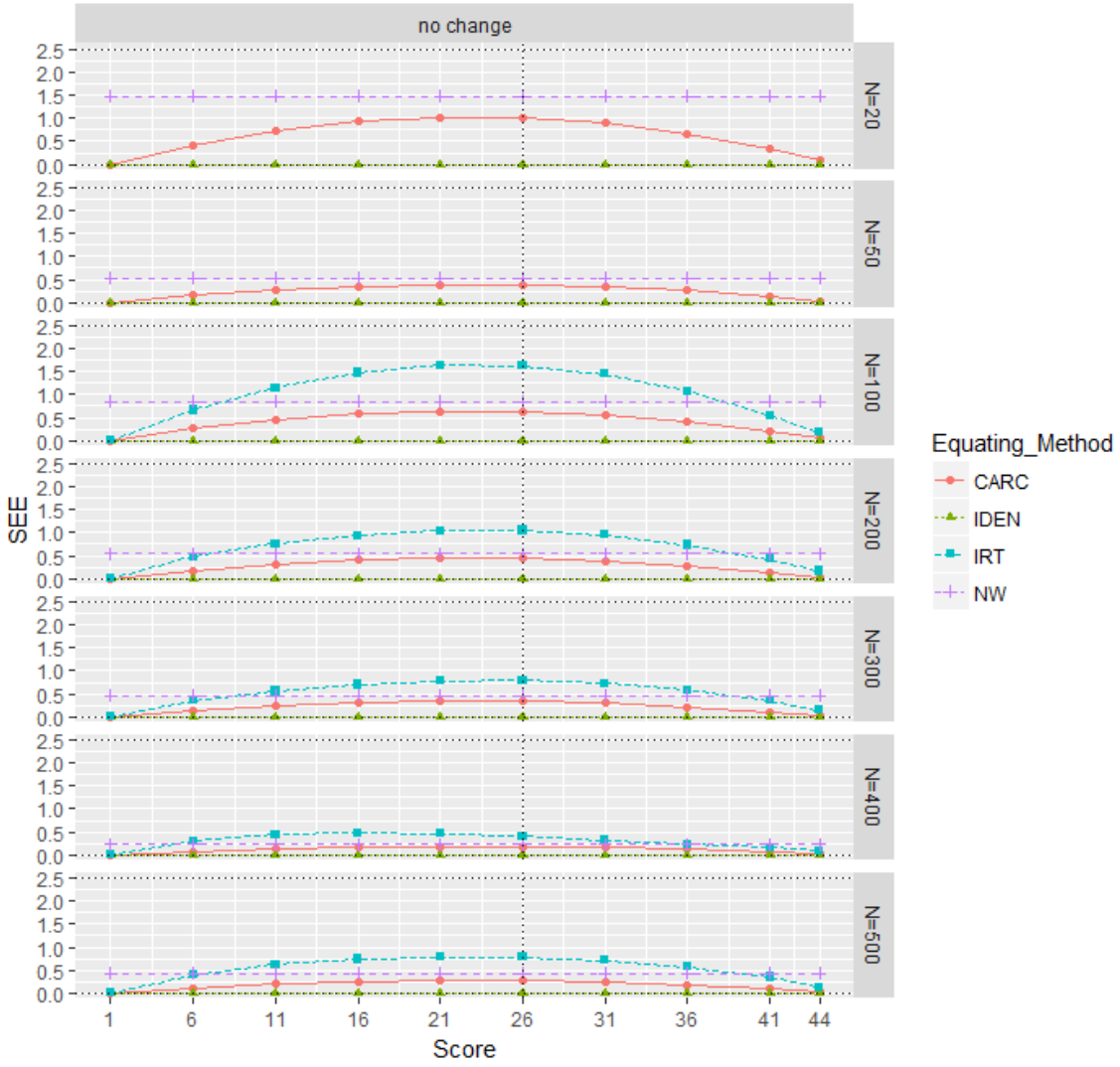
4.3 Effects on Conditional SEE

The current section describes the conditional standard error resulted from the different test and equating conditions. Similar to the above section, the first subsection focuses on the CSEE when equating with non-problematic anchor while the second section analyzes the equating errors when problematic anchor test was included. Figure 4.10 to Figure 4.14 focus on examining how different small equating techniques impact standard error by holding the repeater distribution with mean of -1.5. Figure 4.15 investigate if conditional errors differed among three different repeater distributions using circle-arc equating.

4.3.1 Non-problematic Anchor

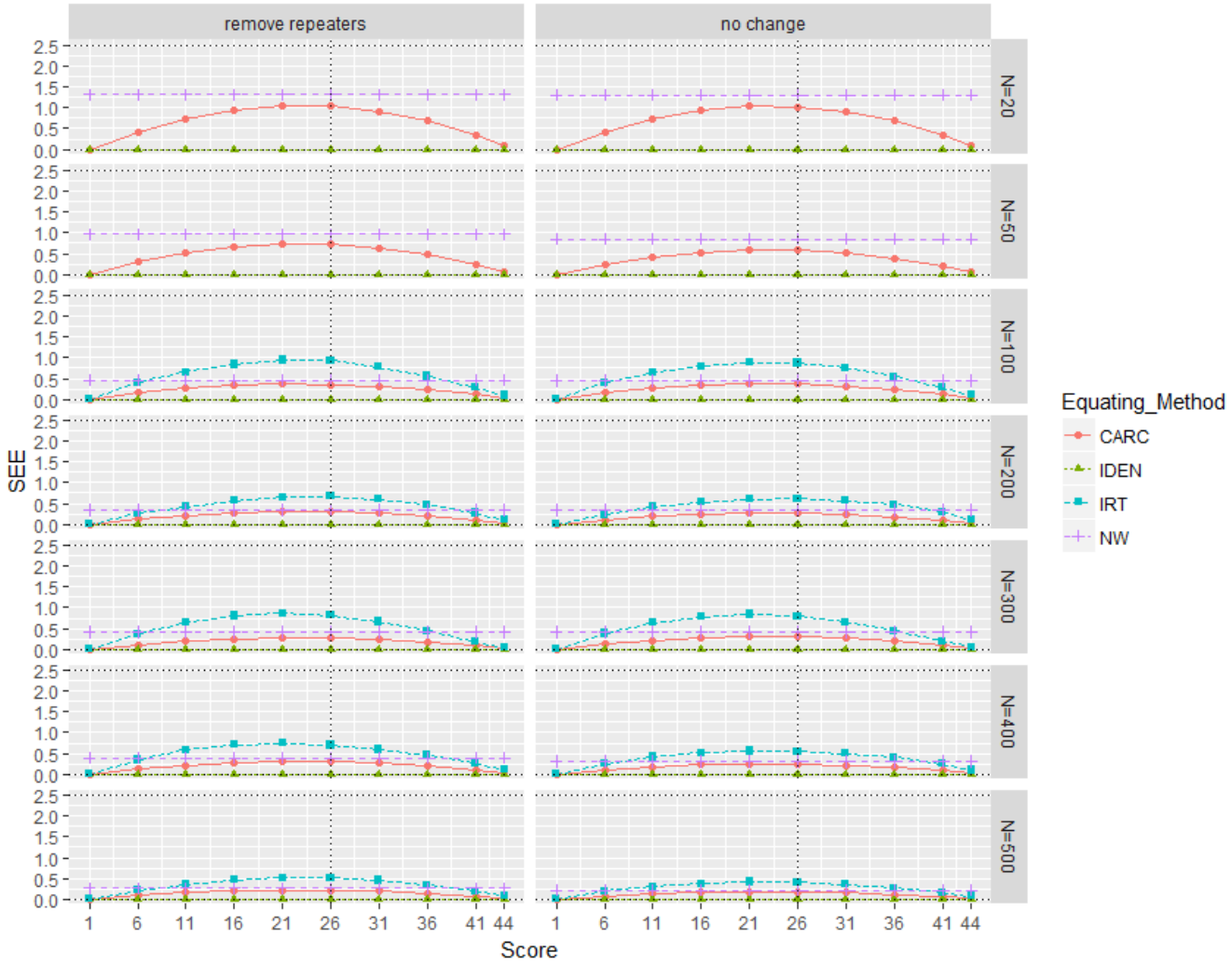
The most striking feature of Figure 4.10 – Figure 4.12 is that the CSEE produced by nonlinear equating methods (i.e., circle-arc and Rasch equating) displayed a nonlinear curve pattern across score ranges. The values of CSEE was the highest near cut-score point and decreased to zero at two ends of score range. The CSEE produced by identity equating was equal to zero across score range while nominal weight mean equating produced a constant CSEE value across score scale. In Figure 4.10 to Figure 4.12, nominal weight equating yields CSEE close to or larger than circle-arc equating. This finding is consistent across all sample size levels. Rasch equating produced largest CSEE in the middle score point; however, if the scores were beyond the two intersection points between Rasch equating and nominal weight equating, nominal weight mean equating produced slightly higher conditional error than Rasch equating. Regarding the factor with the strongest influence on CSEE, sample size had a direct effect in decreasing equating

standard errors. Apparently, the CSEE decreased as sample size increased from 20 to 500. This observation was true for all equating techniques (except identity equating), repeater effect solutions and repeater proportion conditions. The pattern regarding repeater effect solutions was consistent between Figure 4.11 and Figure 4.12. Removing repeater can produce larger standard errors than retaining all examinees. The difference between removing repeaters and retaining repeaters in standard equating errors was more apparent under 35% repeater condition, particularly for size level below 200.



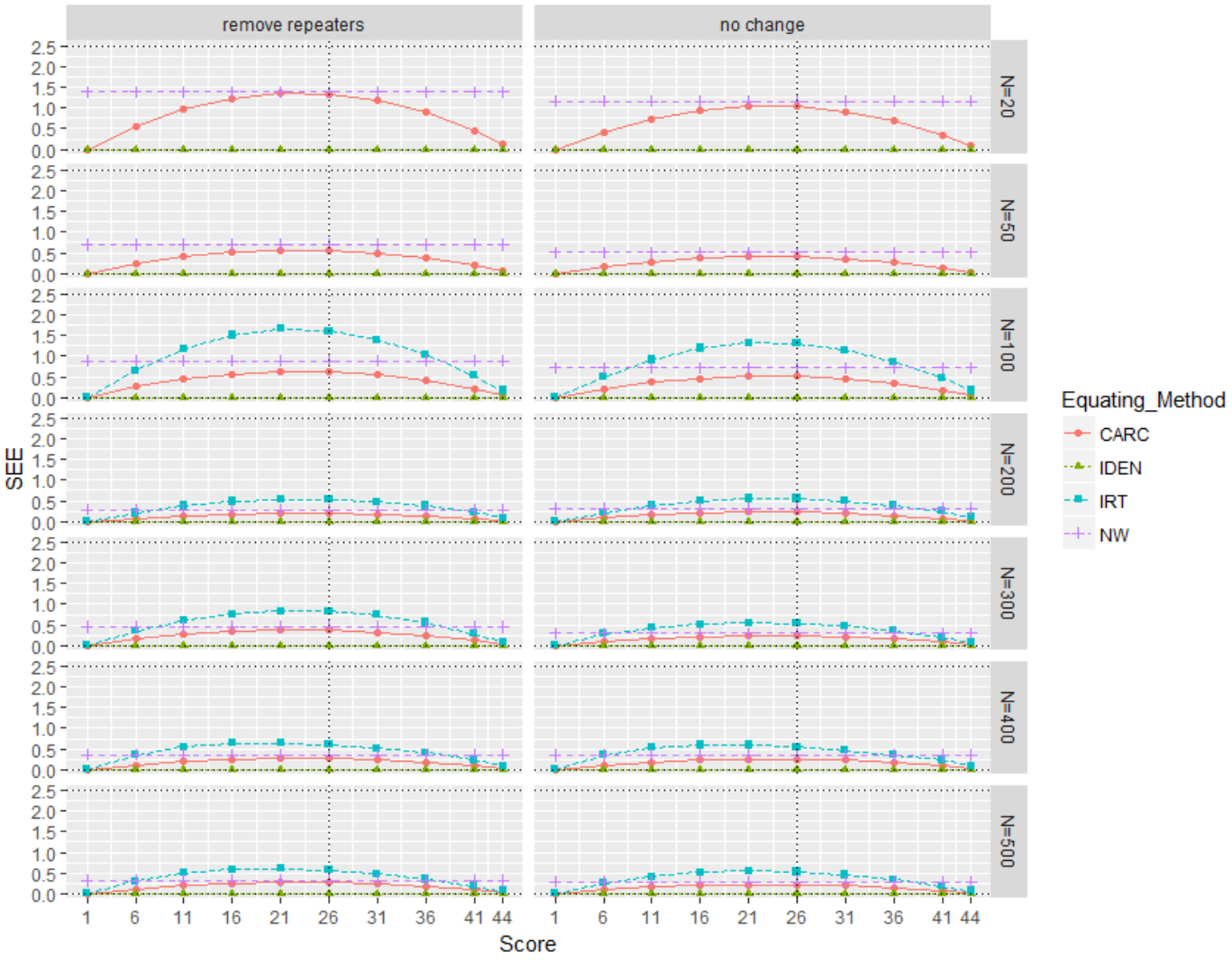
Note. Repeaters follow a distribution $\theta_{R3} \sim N$ (-1.5, 1)

Figure 4.10. SEE Non-problematic Anchor Test of 0% Repeaters by Equating Methods



Note. Repeaters follow a distribution $\theta_{R3} \sim N$ (-1.5, 1)

Figure 4.11. SEE of Non-problematic Anchor Test with 25% Repeaters by Equating Methods

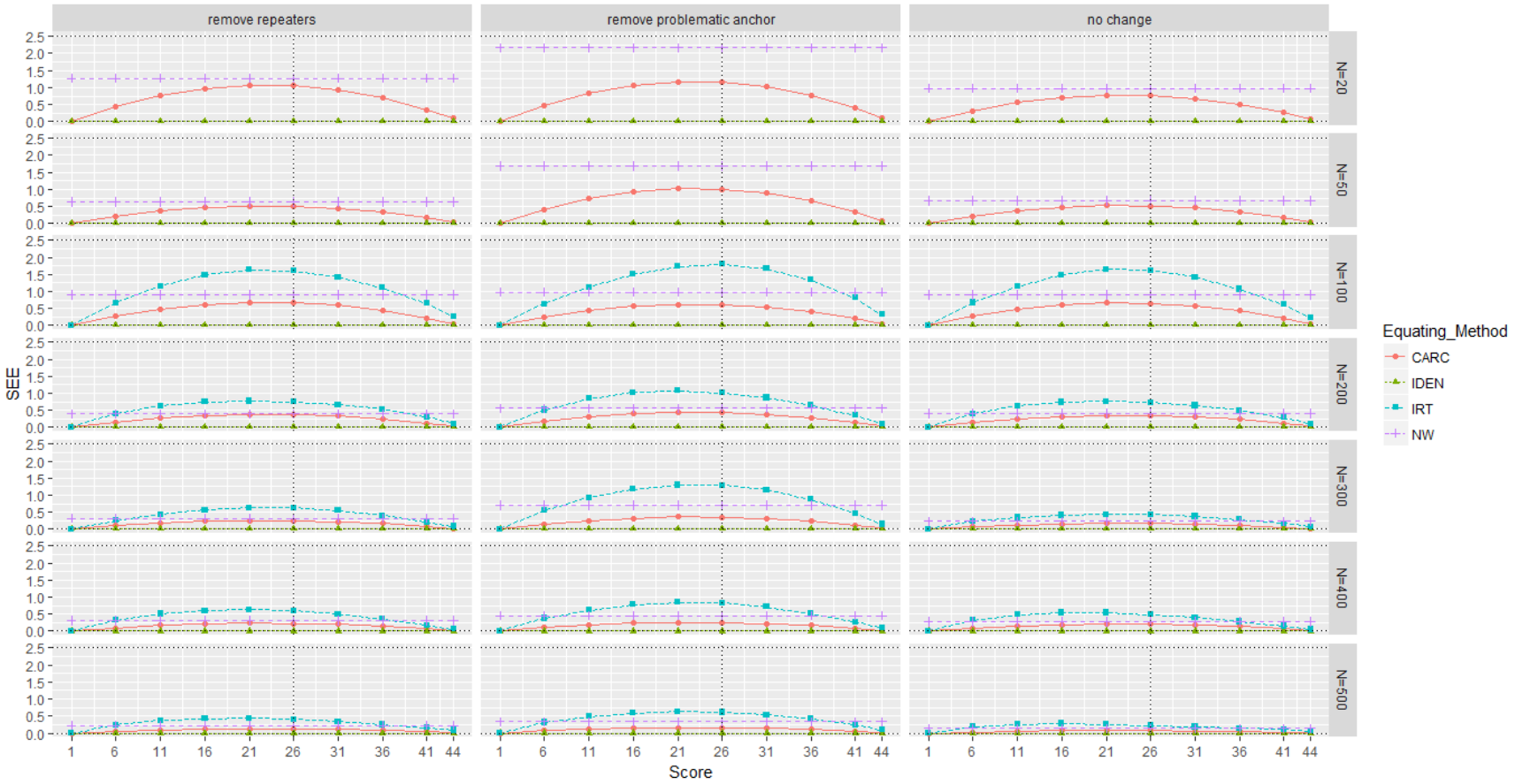


Note. Repeaters follow a distribution $\theta_{R3} \sim N$ (-1.5, 1)

Figure 4.12. SEE of Non-problematic Anchor Test with 35% Repeaters by Equating Methods

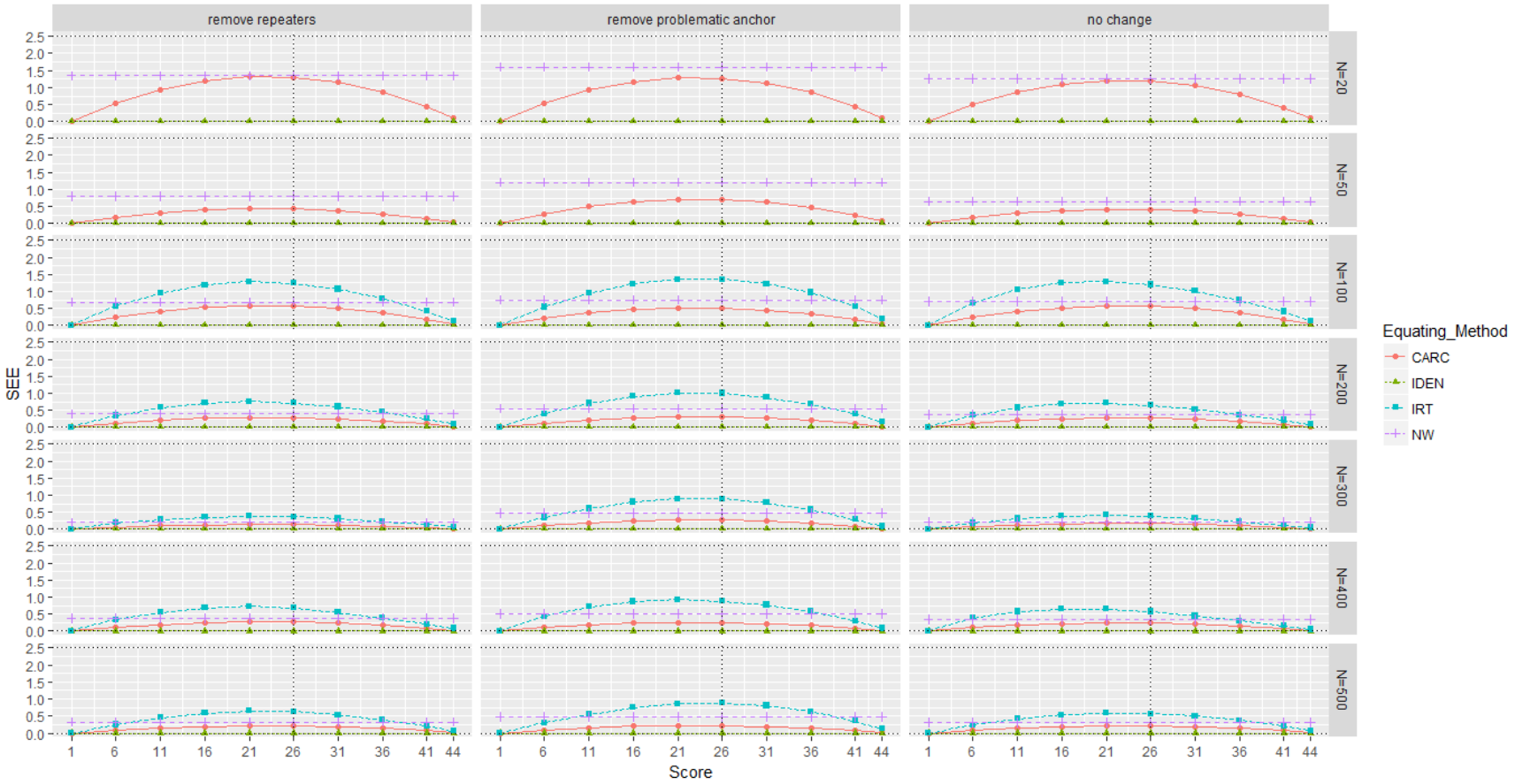
4.3.2 Problematic Anchor

Figure 4.13 and Figure 4.14 portray the CSEE when equating with a problematic anchor. The patterns of CSEE have some similarities between two repeater proportion conditions. First, both figures show a clear pattern that nominal weight mean equating produced higher standard errors than circle-arc equating across score range. The Rasch equating yielded higher CSEE than nominal weight mean equating, especially near the cut-score point. However, nominal weight mean equating produced slight higher SEE at upper and lower score points. Within two intersection points, the order of equating techniques that provided decreasing accuracy is: Identity equating, circle-arc equating, nominal weight equating, and Rasch Equating. If the scores were beyond the intersection points between Rasch and nominal weight mean equating, the sequence would be: Identity equating, circle-arc equating, Rasch Equating, and nominal weight mean equating. Similar to non-problematic anchor condition, smaller sample size decreased equating accuracy. This find applied to all data management strategies. In general, the overall patterns in CSEE between problematic and non-problematic anchor conditions are consistent. In the next sections, the overall equating accuracy across score points would be fully described based on the values of WSEE.



Note. Distribution of repeaters: $\theta_{R3} \sim N$ (-1.5, 1)

Figure 4.13. SEE of Problematic Anchor Test with 25% Repeaters by Equating Methods

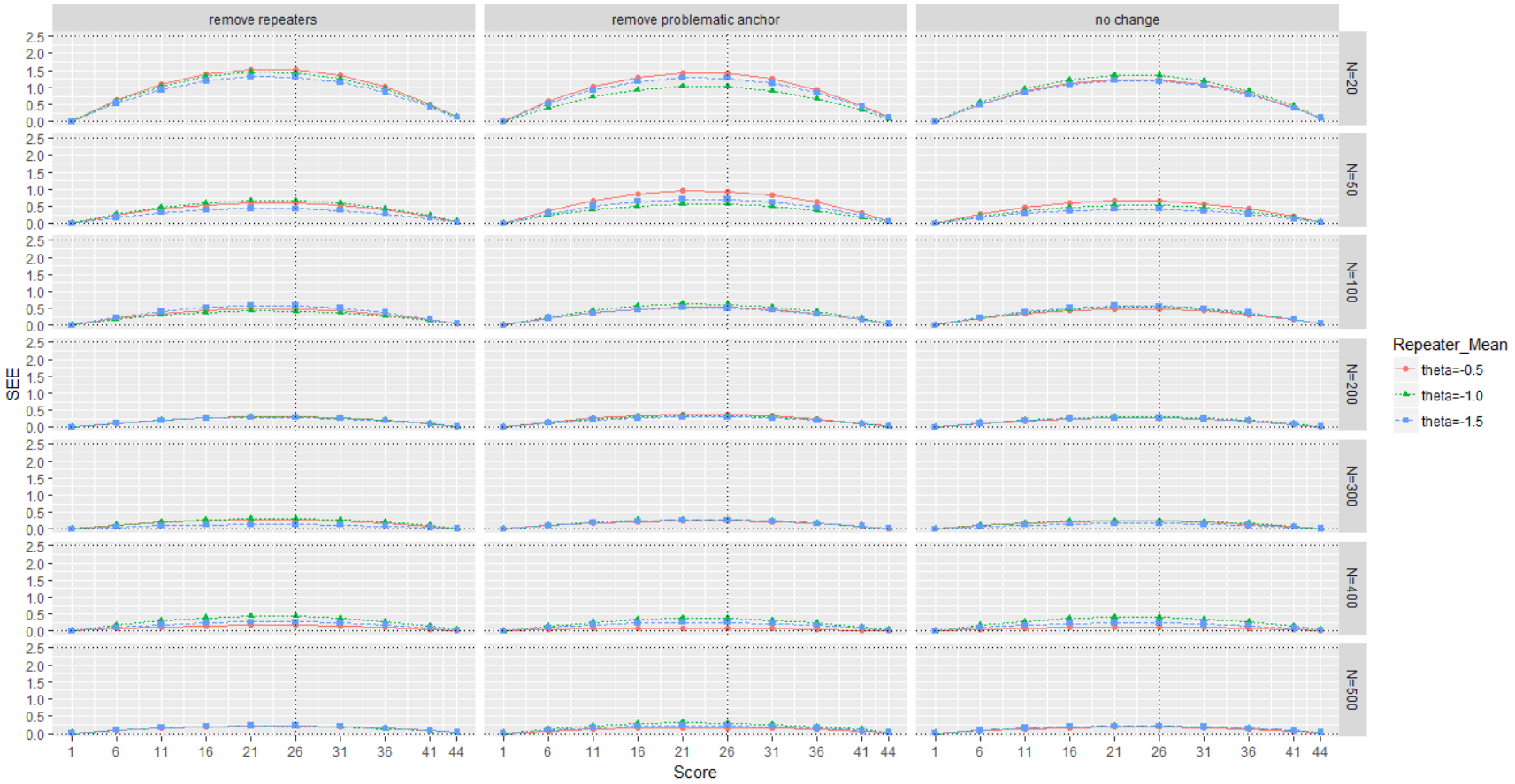


Note. Distribution of repeaters: $\theta_{R3} \sim N$ (-1.5, 1)

Figure 4.14. SEE of Problematic Anchor Test with 35% Repeaters by Equating Methods

4.3.3 Repeater Mean

Unsurprisingly, increasing sample size levels have the most pronounced direct effect in reducing CSEE. Figure 4.15 indicates that the lines representing different repeater means are close across all conditions. Although there was a trivial difference between $\theta_{R1} \sim N$ (-0.5, 1) and $\theta_{R2} \sim N$ (-1.0, 1) distributions under "removing problematic anchor" solution. This mainly occurred under small sample size levels in which the variation of CSEE was likely to be larger than other sample size levels.



Note. Equating Method: Circle-Arc Equating

Figure 4.15. SEE of Problematic Anchor Test with 35% Repeaters by Repeater Mean

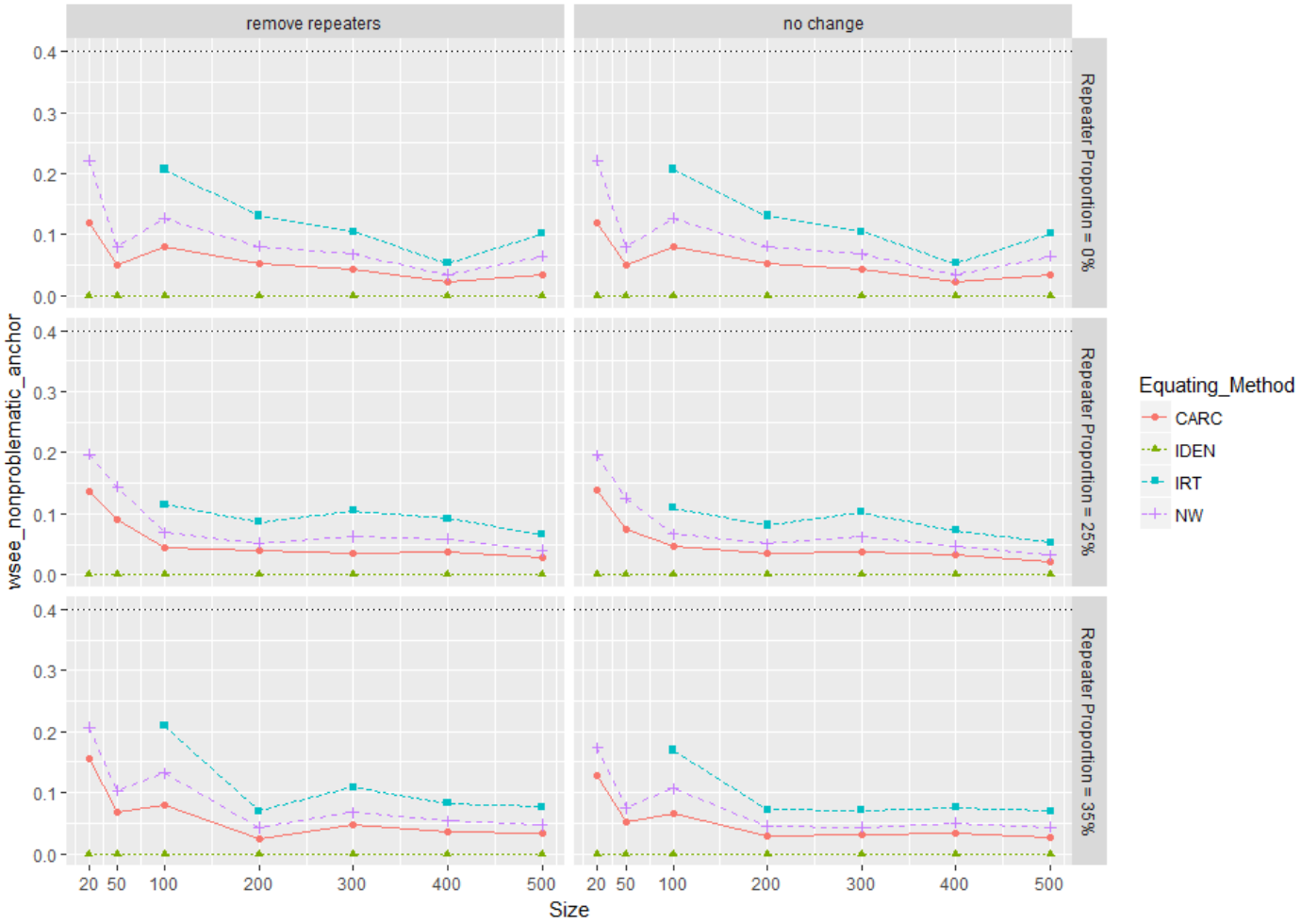
4.4 Effect on WSEE

WSEE indicates the overall equating errors across score points by considering the ratio of examinees at each score point. The current section consists of two subsections. The first subsection focuses on analyzing overall equating error under non-problematic anchor equating condition. The second subsection investigates WSEE patterns under an equating condition with a problematic anchor. The WSEE are described in figures and numerical approach. Figure 4.16 and Figure 4.17 emphasize how WSEE differ between anchor test conditions while Figure 4.18 is an example showing if the WSEE values are different between three repeater distributions. In addition, each subsection reports the mean and standard deviation by sample size levels, repeater effect solutions, equating techniques, proportion of repeaters and repeater mean in Table B8 and Table B9.

4.4.1 Non-problematic Anchor

Figure 4.16 has 6 charts, the charts in the same row refer to the repeater proportion condition and the column refers to the repeater effects solutions. The charts in the first column show the WSEE under “removing repeaters” solution and the charts at the second column show the WSEE under “retaining repeaters” condition. This figure only has 6 charts because equating was performed with non-problematic anchor and there is no need to display “removing problematic anchor” solution. Visual inspection of Figure 4.16 reveals that the Rasch equating produced the largest overall SEE. In terms of the mean value, the sequence of equating techniques providing decreasing accuracy is: identity equating (M = 0.00, SD = 0.00), circle-arc equating (M = 0.06, SD = 0.04), nominal weight mean equating (M = 0.10, SD = 0.06) and Rasch equating (M = 0.11 SD

= 0.04). The mean of WRMSB across sample size levels ranged from 0.04 (*N* = 500, *SD* = 0.03) to 0.12 (*N = 20, SD* = 0.09). In the figure, there is a clear pattern that WSEE decreased as sample size increased and the elbow located at the sample size level *N* = 200. The means of WSEE under "removing repeater" and "retaining repeater" condition were 0.06 (*SD* = 0.06) and 0.06 (*SD* = 0.05), respectively. The figure also confirms that WSEE patterns were similar between two data management conditions. According to summary statistics, the proportion of repeaters did not substantially impact WSEE. The mean and standard deviation for 0%, 25% and 35% condition were 0.06 (*SD* = 0.06), 0.06 (*SD* = 0.06), 0.06 (*SD* = 0.06), respectively.



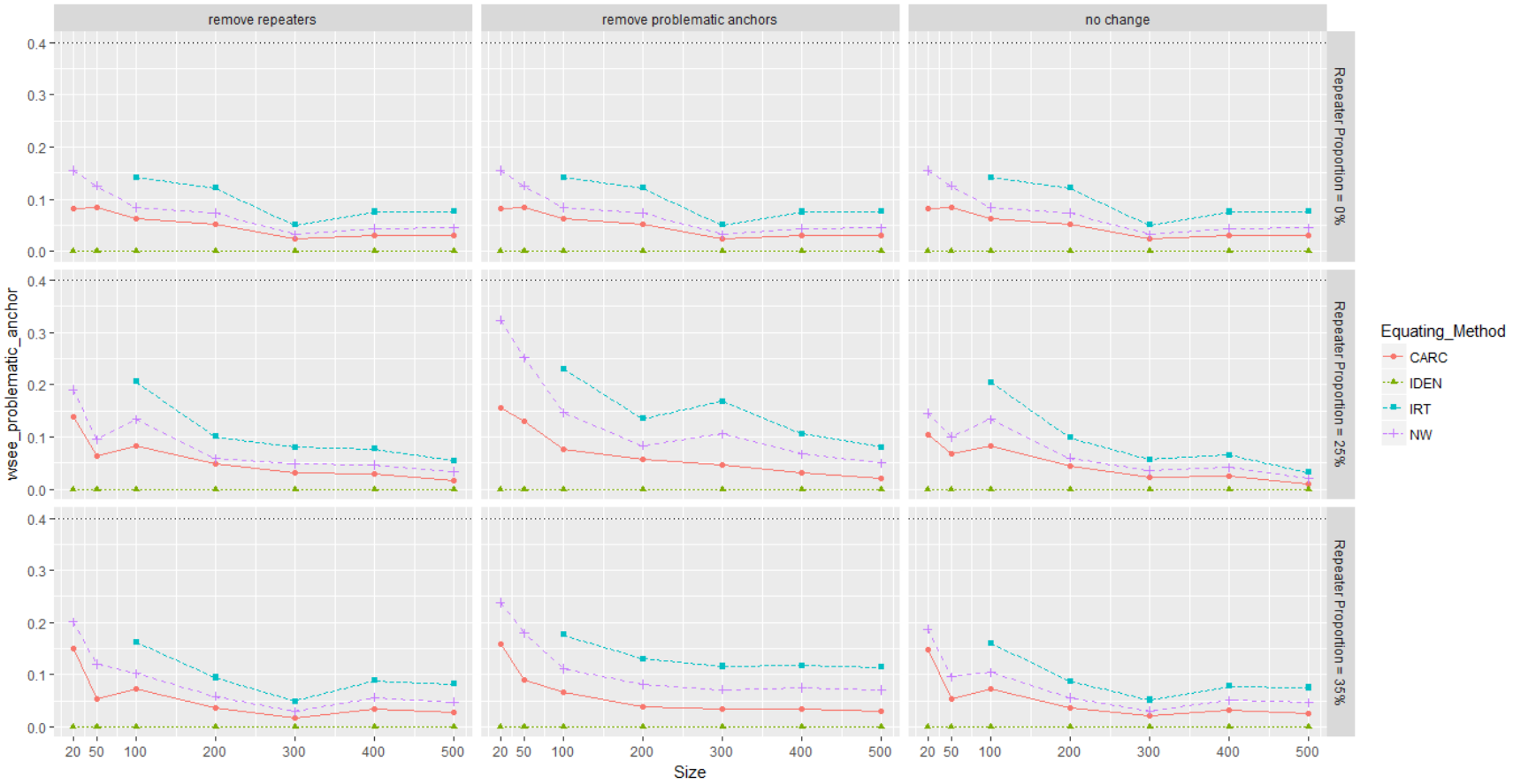
Note. Distribution of repeaters: $\theta_{R3} \sim N$ (-1.5, 1)

Figure 4.16. WSEE of Non-problematic Anchor Test by Equating Methed

4.4.2 Problematic Anchor

Figure 4.17 shows values of WSEE when equating with drifted anchor items. The charts at the first column show the WSEE under "removing repeaters" condition, the second column displays the charts of "removing problematic anchor" solution and the graphs at the third column depict the SEE under "retaining repeaters and problematic anchor" condition. The magnitude of WSEE was not remarkably different between "removing repeaters condition" (M = 0.06, SD = 0.06), "excluding problematic anchor" (M = 0.07, SD = 0.07) and "retaining all items and repeaters" (M = 0.06, SD = 0.06) conditions. However, the disparities in WSEE between equating techniques was slightly higher under "removing problematic anchor" condition. A closer examination on the WSEE under this solution shows that nominal weight mean equating had substantially high WSEE at size level of 20 and 50. The value of WSEE produced by nominal weight equating exceeded 0.3 if 20 examinees were included equating procedure. The mean WSEE produced by circle-arc equating, identity equating, Rasch equating and nominal weight mean equating were 0.06 (SD = 0.04), 0.00 (SD = 0.00), 0.11 (SD = 0.04) and 0.10 (SD = 0.06), respectively. The graphs also show that Rasch equating and nominal weight mean equating consistently produced larger error than circle-arc equating regardless of the presence of problematic anchor items. The mean of WSEE across sample size levels ranged from 0.03 (SD = 0.04) at N = 500 sample size level to 0.13 (SD = 0.11) at N =20, which was similar to non-problematic anchor equating condition. In sum, sample size and equating techniques were two factors impacting WSEE. The overall means for non-problematic and problematic anchor equating condition were 0.06 (SD = 0.06) and 0.06 (SD = 0.07), respectively. This may indicate drifted anchor had very little

effect in WSEE. The patterns of WSEE did not substantially differ across repeater proportion conditions and repeater effect solutions.

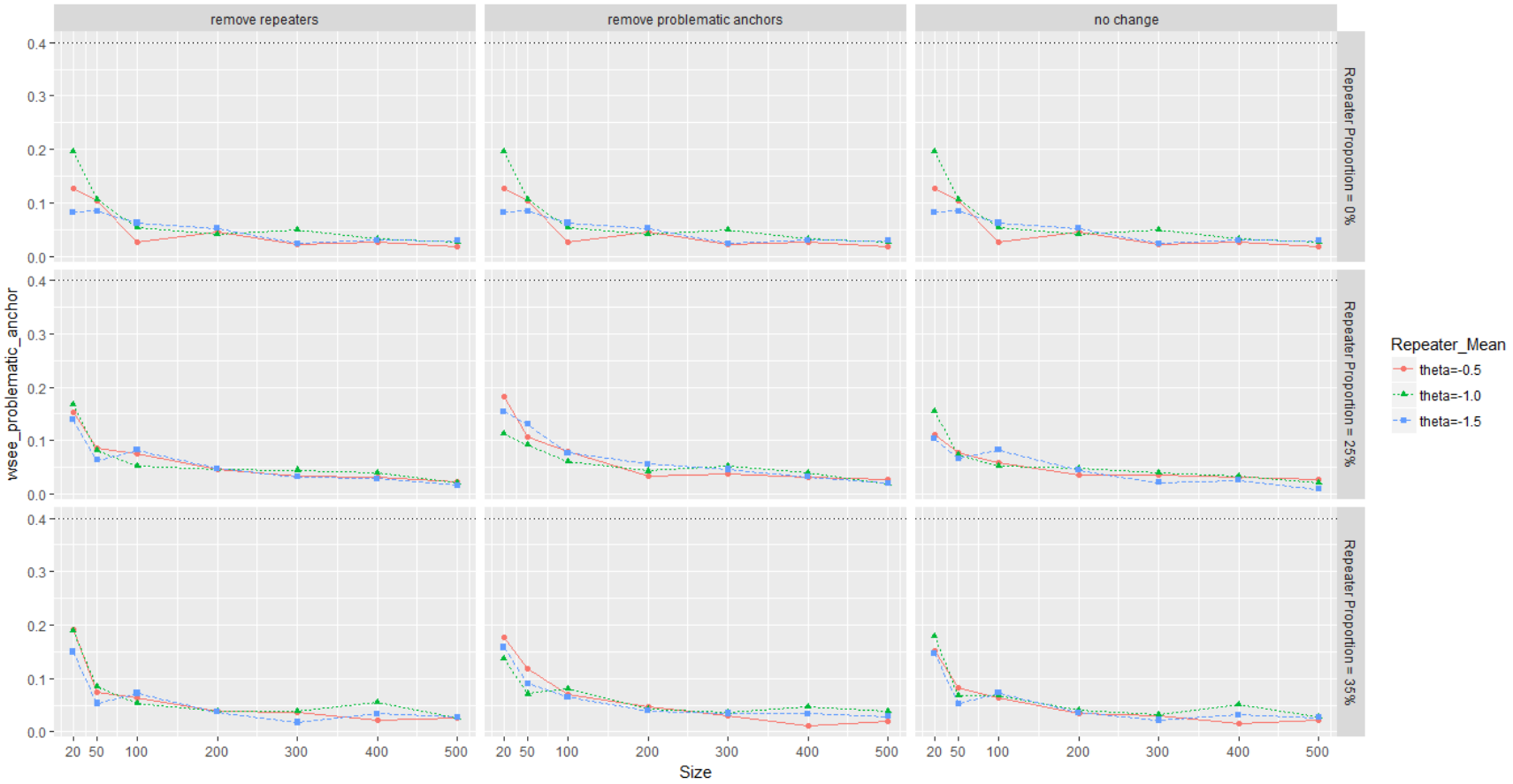


Note. Distribution of repeaters: $\theta_{R3} \sim N$ (-1.5, 1)

Figure 4.17. WSEE of Problematic Anchor Test by Equating Method

4.4.3 Repeater Mean

For non-problematic anchor condition, the mean and standard deviation for $\theta_{R1} \sim N$ (-0.5, 1), $\theta_{R2} \sim N$ (-1.0, 1) and $\theta_{R3} \sim N$ (-1.5, 1) were 0.06 (*SD* = 0.06), 0.06 (*SD* = 0.05), 0.06 (*SD* = 0.06), respectively. If the anchor test was drifted because of repeaters, the mean and standard deviation for $\theta_{R1} \sim N$ (-0.5, 1), $\theta_{R2} \sim N$ (-1.0, 1) and $\theta_{R3} \sim N$ (-1.5, 1) were 0.06 (*SD* = 0.07), 0.07 (*SD* = 0.07), 0.06 (*SD* = 0.06), respectively. In Figure 4.18, lines representing different repeater mean lie over each other across sample size levels, confirming the small influence of repeater mean.



Note. Equating Method: Circle-Arc Equating

Figure 4.18. WSEE of Problematic Anchor Test by Repeater Mean

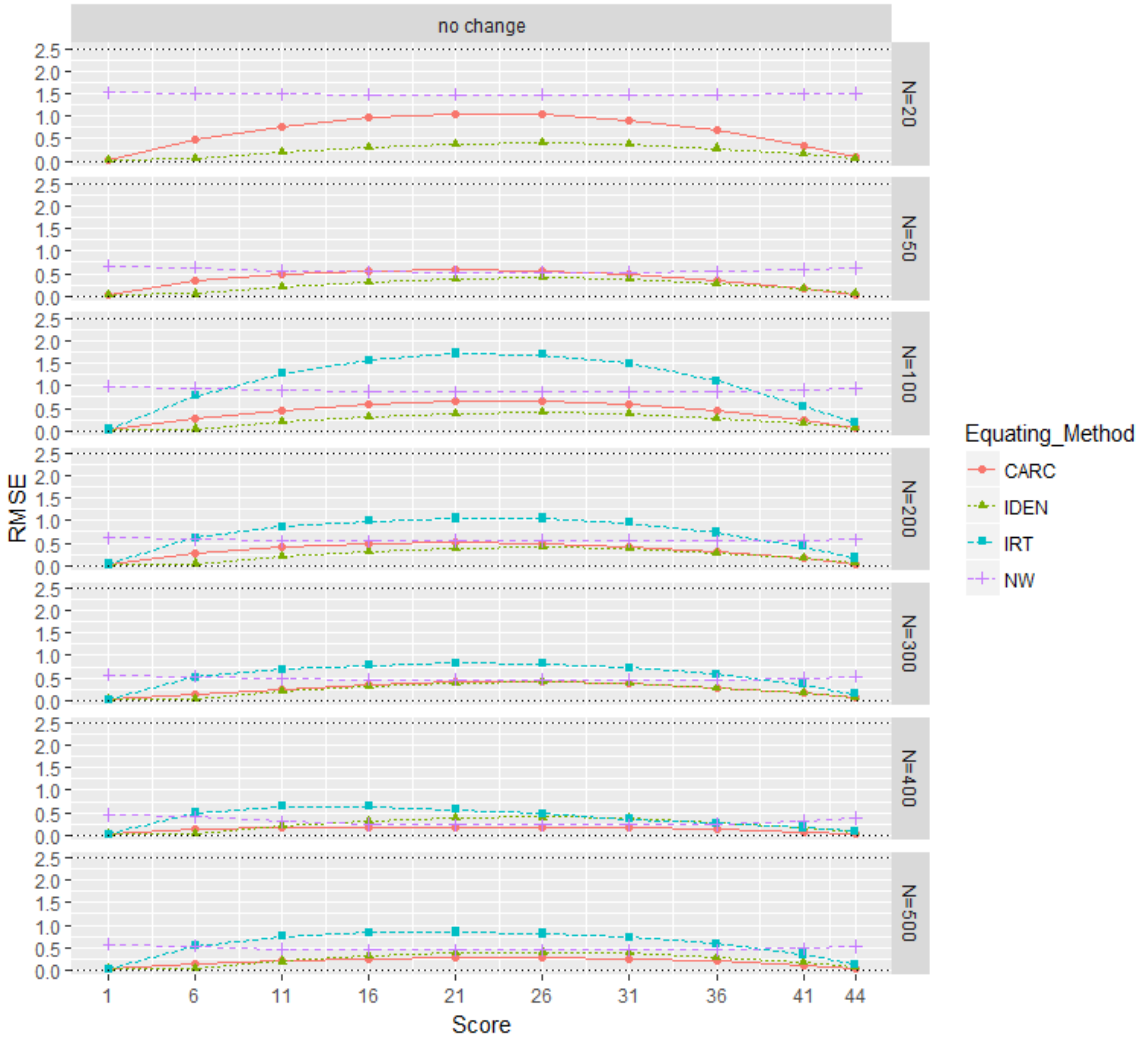
4.5 Effects on Conditional RMSE

RMSE of equating is a single statistic reflects the combination of random error and systematic error. The patterns of RMSE synthesizing the results of bias and SEE. Figure 4.19 depicts the patterns of RMSE across scale when equating with no problematic anchors or repeaters. The RMSE charts presented in this figure is considered as a "baseline" RMSE because no repeaters were included in the dataset. Figure 4.20 and Figure 4.21 depict the RMSE under 25% repeaters and 35% repeaters conditions when equating with non-problematic anchors. Figure 4.22 and Figure 4.23 display the conditional RMSE resulted from problematic anchor equating conditions with 25% and 35% repeaters in the new form. Figure 4.24 addresses how conditional RMSE patterns differed if the repeater distributions had different mean values.

4.5.1 Non-problematic Anchor

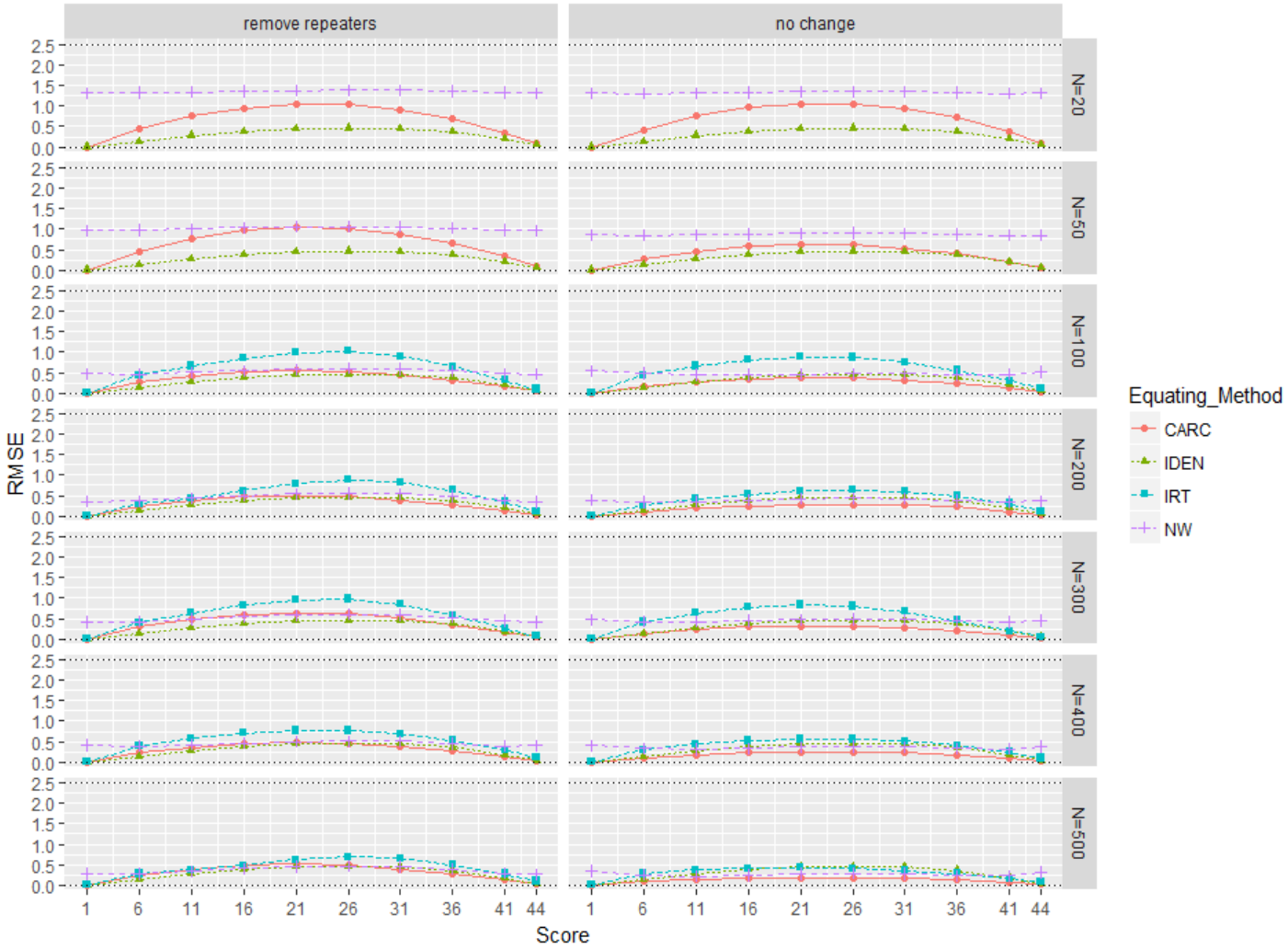
Figure 4.19 - Figure 4.21 display the RMSE patterns across score scale as the proportion of repeater increase from 0% to 35%. The patterns of RMSE are consistent across repeater proportion levels. In each proportion, the magnitude of RMSE provided by nominal weight mean equating was constant across score points while Rasch equating, circle-arc equating and identity equating produced larger RMSE in the middle of score scale and smaller RMSE at upper and lower ends of score scale. Nominal weight mean equating produced larger RMSE than circle-arc equating; circle-arc equating produced greater RMSE than identity equating. Similar to the patterns of CSEE, the lower and upper intersection score points between nominal mean equating and Rasch equating were close to 6 and 36 points. Within the intersection points interval, the sequence of equating

techniques for decreasing RMSE was: Rasch equating, nominal weight mean, circle-arc and identity equating. As sample size increased, the overall RMSE decreased and the gaps between classical equating techniques were noticeable if $N < 200$. Under the same proportion condition, removing repeaters produced slightly higher RMSE than retaining repeaters. The interaction between data management approach and proportion would be fully examined based on overall RMSE, which is WRMSE, in the following section.



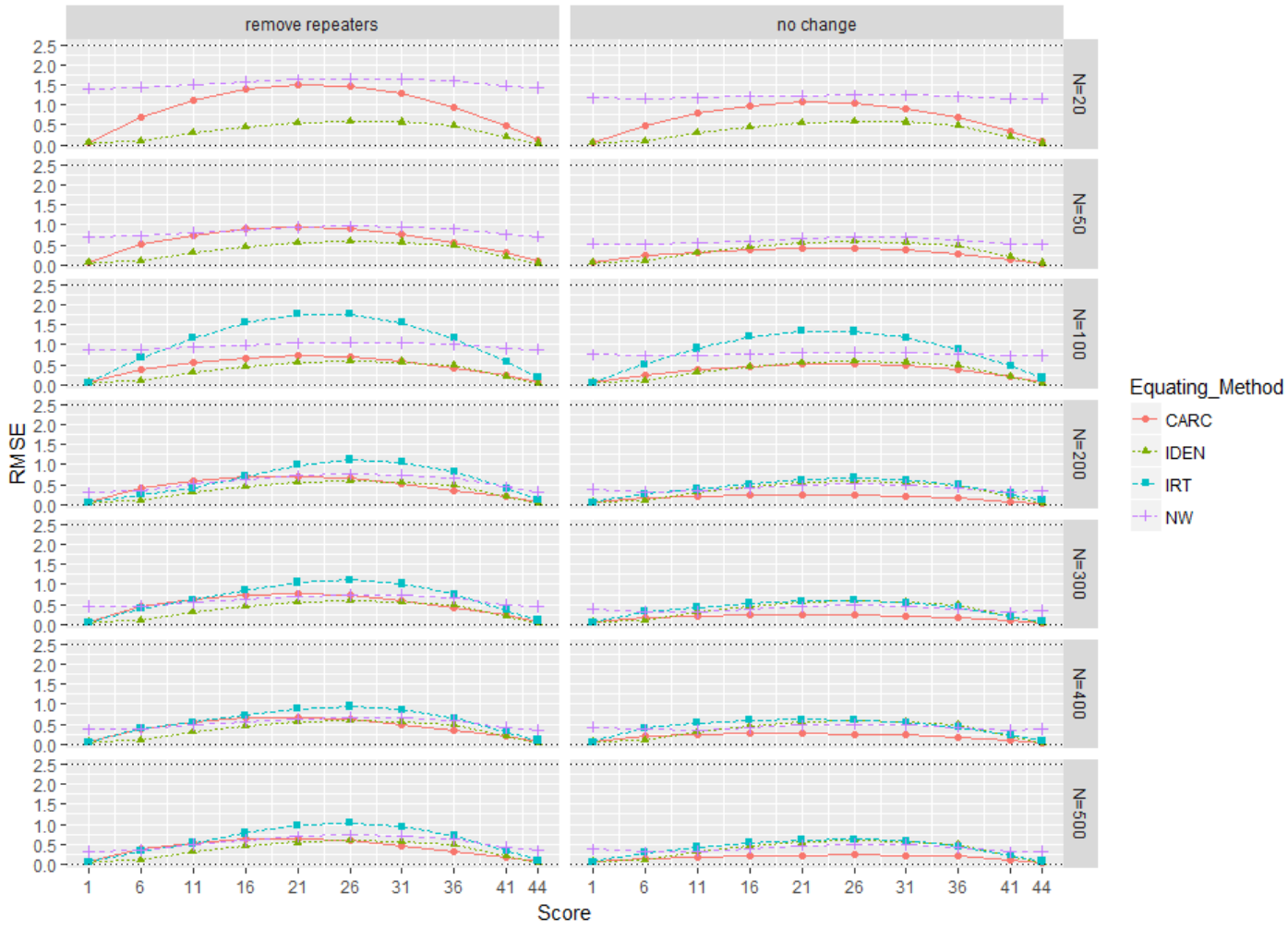
Note. Distribution of repeaters: $\theta_{R3} \sim N$ (-1.5, 1)

Figure 4.19. RMSE of Non-problematic Anchor with 0% Repeaters by Equating Methods



Note. Distribution of repeaters: $\theta_{R3} \sim N$ (-1.5, 1)

Figure 4.20. RMSE of Non-problematic Anchor Test with 25% Repeaters by Equating Methods



Note. Distribution of repeaters: $\theta_{R3} \sim N$ (-1.5, 1)

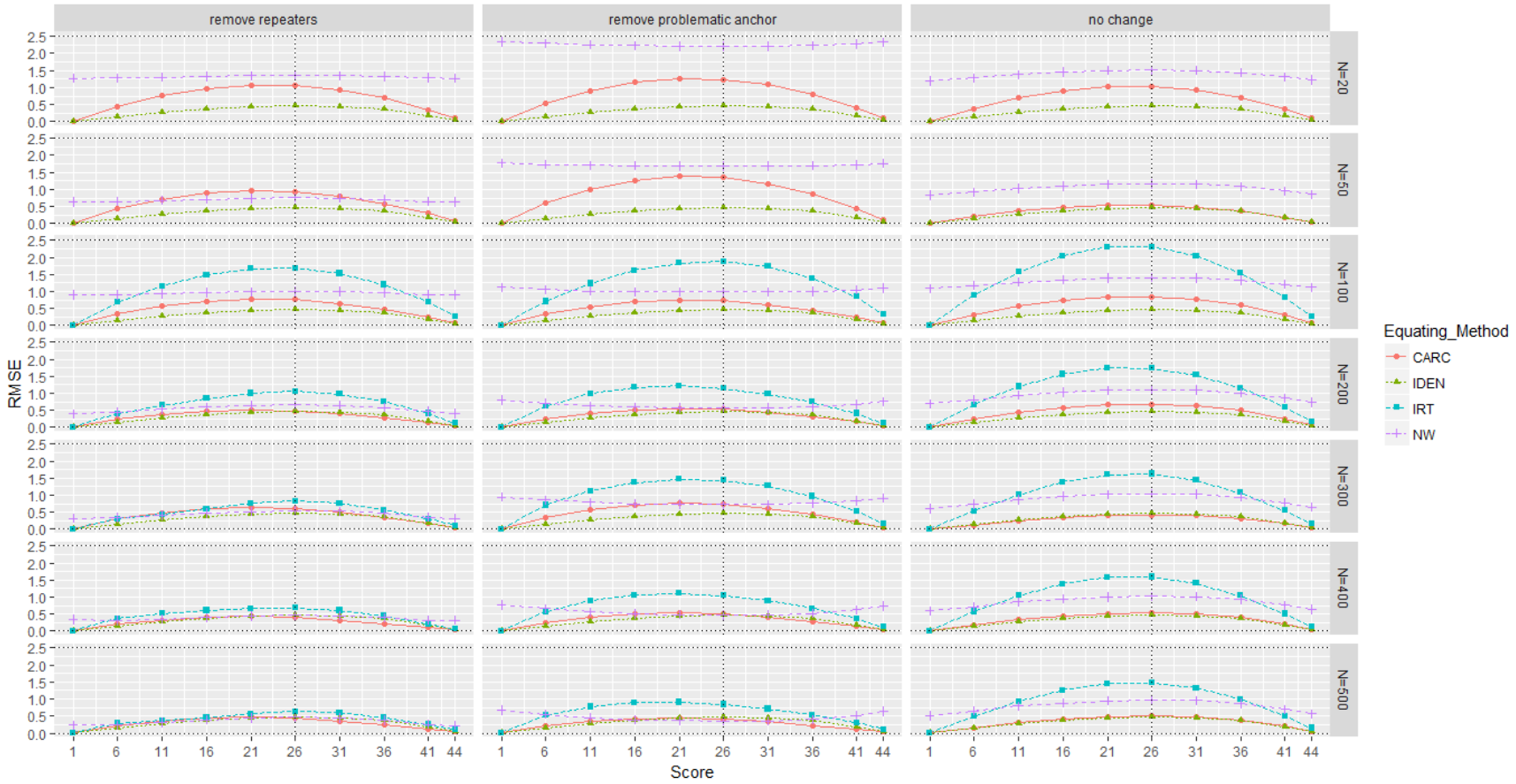
Figure 4.21. RMSE of Non-problematic Anchor Test with 35% Repeaters by Equating Methods

4.5.2 Problematic Anchor

In Figure 4.22 and Figure 4.23, larger RMSE was found in the middle of score scale and gradually decreased to zero toward the two-ends of the score scale for most of equating techniques. This pattern applied to circle-arc, identity and Rasch equating. The patterns of nominal weight mean equating produced a relatively constant RMSE across score scale. Similar to the RMSE resulted from the non-problematic equating condition, nominal weight mean equating produced the highest RMSE across raw score scales among three classical equating techniques. Circle-arc equating yielded larger RMSE than identity equating but likely to lie over identity equating as the sample size increased to 200. Rasch equating produced the highest RMSE except for the upper and lower ends of the scale. Under 25% repeater condition, increasing sample size can decrease the RMSE; however, the RMSE produced by identity equating was not drastically changed across sample size levels. Under 35% repeater condition, RMSE produced by circle-arc equating progressively decreased as sample size increased from $N = 20$ to $N = 100$. The influence in terms of sample size was not notable for all equating methods if $N \geq 200$. According to Figure 4.22 and Figure 4.23, retaining repeaters and problematic anchor were likely to provide the highest level of RMSE than applying repeater effect solutions. This finding was particularly striking for nominal weight mean and Rasch equating techniques.

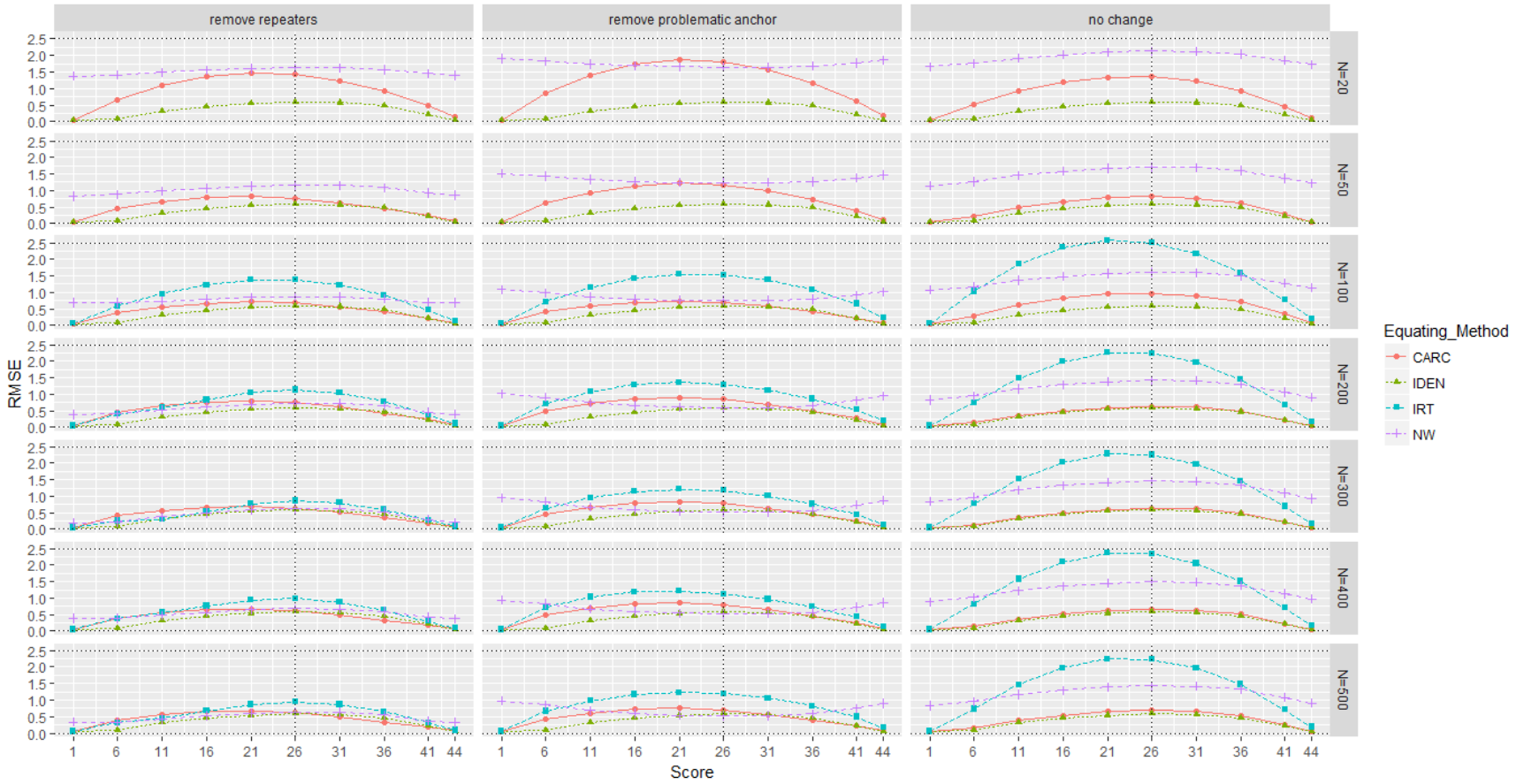
Compared with the non-problematic equating condition, the problematic equating condition created larger RMSE over the entire score scale, particularly for "retaining all repeaters and anchors" conditions. In addition, Rasch equating tended to create larger RMSE by holding sample size, repeater proportion and repeater effect solution constant. Under the sample size levels where Rasch equating was absent ($N > 100$), nominal

weight mean equating and circle-arc equating resulted in larger RMSE than identity equating.



Note. Distribution of repeaters: $\theta_{R3} \sim N$ (-1.5, 1)

Figure 4.22. RMSE of Problematic Anchor Test with 25% Repeaters by Equating Methods

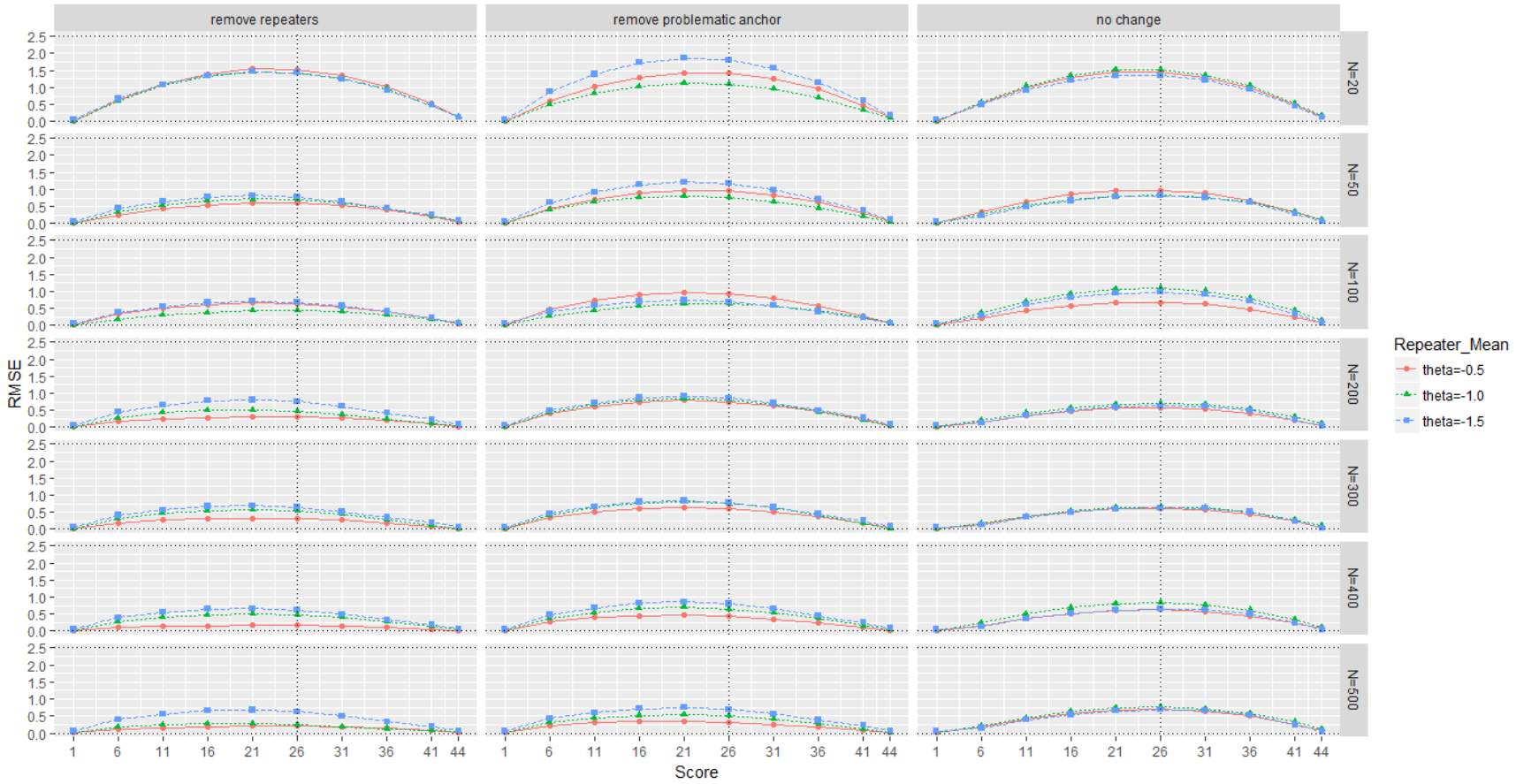


Note. Distribution of repeaters: $\theta_{R3} \sim N$ (-1.5, 1)

Figure 4.23. RMSE of Problematic Anchor Test with 35% Repeaters by Equating Methods

4.5.3 Repeater Mean

Figure 4.24 focuses on investigating if different repeater distributions can lead to different conditional RMSE. The graphs show that lines representing different repeater means are generally close to each other for most conditions. However, small gaps between repeater means are observed from some of the charts. The repeater mean $\theta_{R1} \sim N$ (-0.5, 1) closer to non-repeater was likely to lead lower RMSE especially for size levels larger than 100. As a result, the repeater mean did not substantially impact the conditional RMSE value. The impact of repeater mean was weaker than the influence of sample size.



Note. Equating Method: Circle-Arc Equating

Figure 4.24. RMSE of Problematic Anchor Test with 35% Repeaters by Repeater Mean

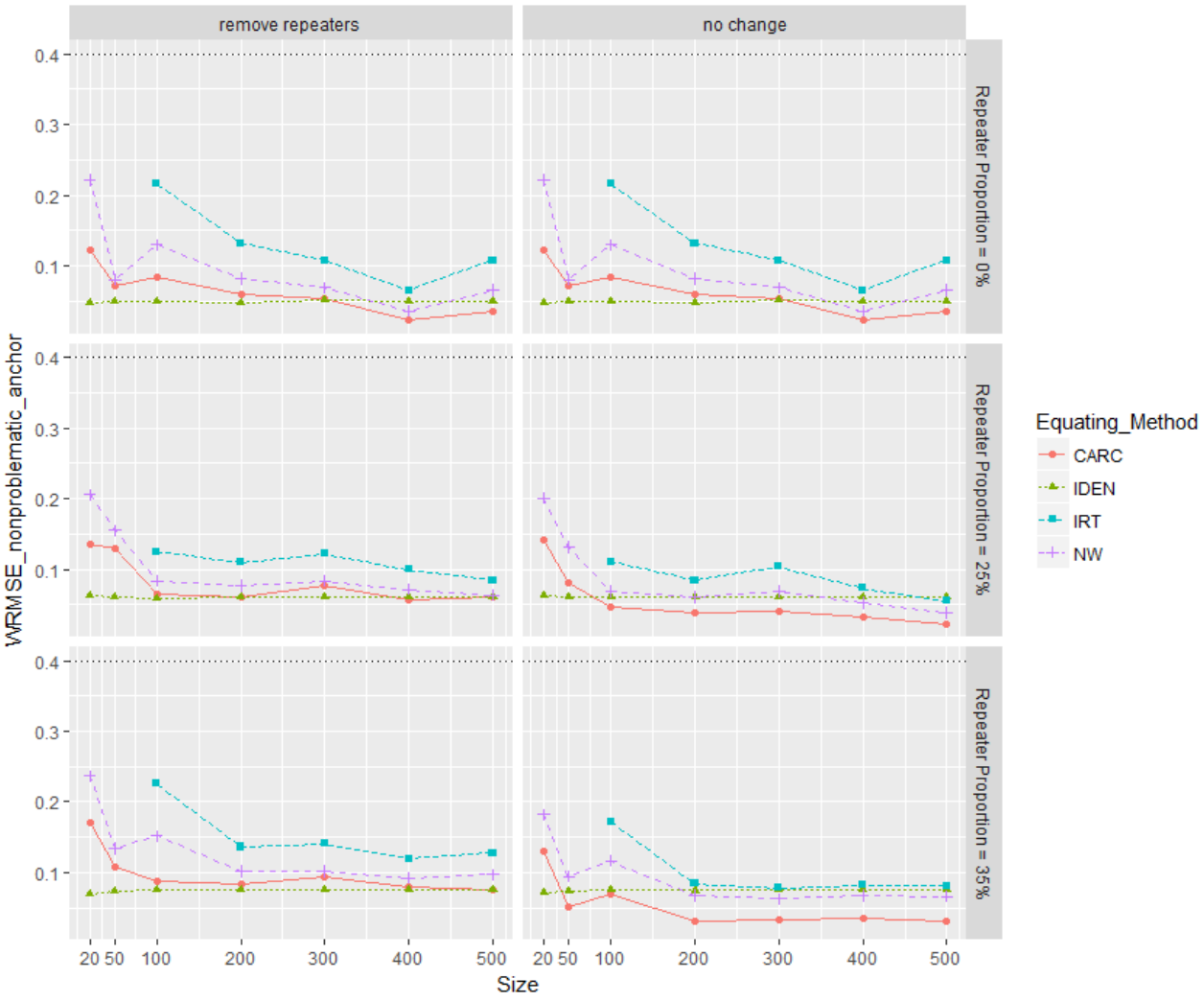
4.6 Effect on WRMSE

WRMSE indicates the combination of WRMSB and WSEE. Figure 4.25 and Figure 4.26 emphasize how WRMSE differ between equating approaches by fixing the repeater mean as $\theta_{R3} \sim N$ (-1.5, 1). Figure 4.27 is an example showing if there are differences between three repeater distributions. The summary statistics are reported in Table B10 and Table B11.

4.6.1 Non-problematic Anchor

There are 2*3 charts in Figure 4.25 where two columns show the WRMSE under "removing repeaters" condition and "retaining repeaters" condition. The first to the third row display condition with 0%, 25%, and 35% repeaters, respectively. The mean of WRMSE across sample size levels ranges from 0.06 (*SD* = 0.02) with size of 500 to 0.14 (*SD* = 0.08) with size of 20. The figure also reveals that the WRMSE at smaller sample size level was larger and had more variability. The performance of equating techniques interacted with data management approaches. The identity equating method was likely to produce a small and stable WRMSE under "removing repeaters" solutions. Circle-arc equating was likely to yield smallest WRMSE under "retaining repeater" condition if repeaters were included in the equating procedure. Nominal weight mean equating provided higher RMSE than circle-arc and identity equating under both data management conditions, the difference was getting larger as more repeaters were added to the total sample. The Rasch equating produced substantial highest WRMSE across all conditions. The mean WRMSE produced by circle-arc equating, identity equating, Rasch equating and nominal weight mean equating were 0.07 (*SD* = 0.04), 0.06 (*SD* = 0.01), 0.11 (*SD* =

0.04) and 0.10 (*SD* = 0.06), respectively. According to summary statistics and graphs, the proportion of repeaters or repeater mean did not remarkably influence WRMSE. The mean and standard deviation for 0%, 25% and 35% conditions were 0.08 (*SD* = 0.05), 0.08 (*SD* = 0.05), 0.09 (*SD* = 0.05), respectively.



Note. Distribution of repeaters: $\theta_{R3} \sim N$ (-1.5, 1)

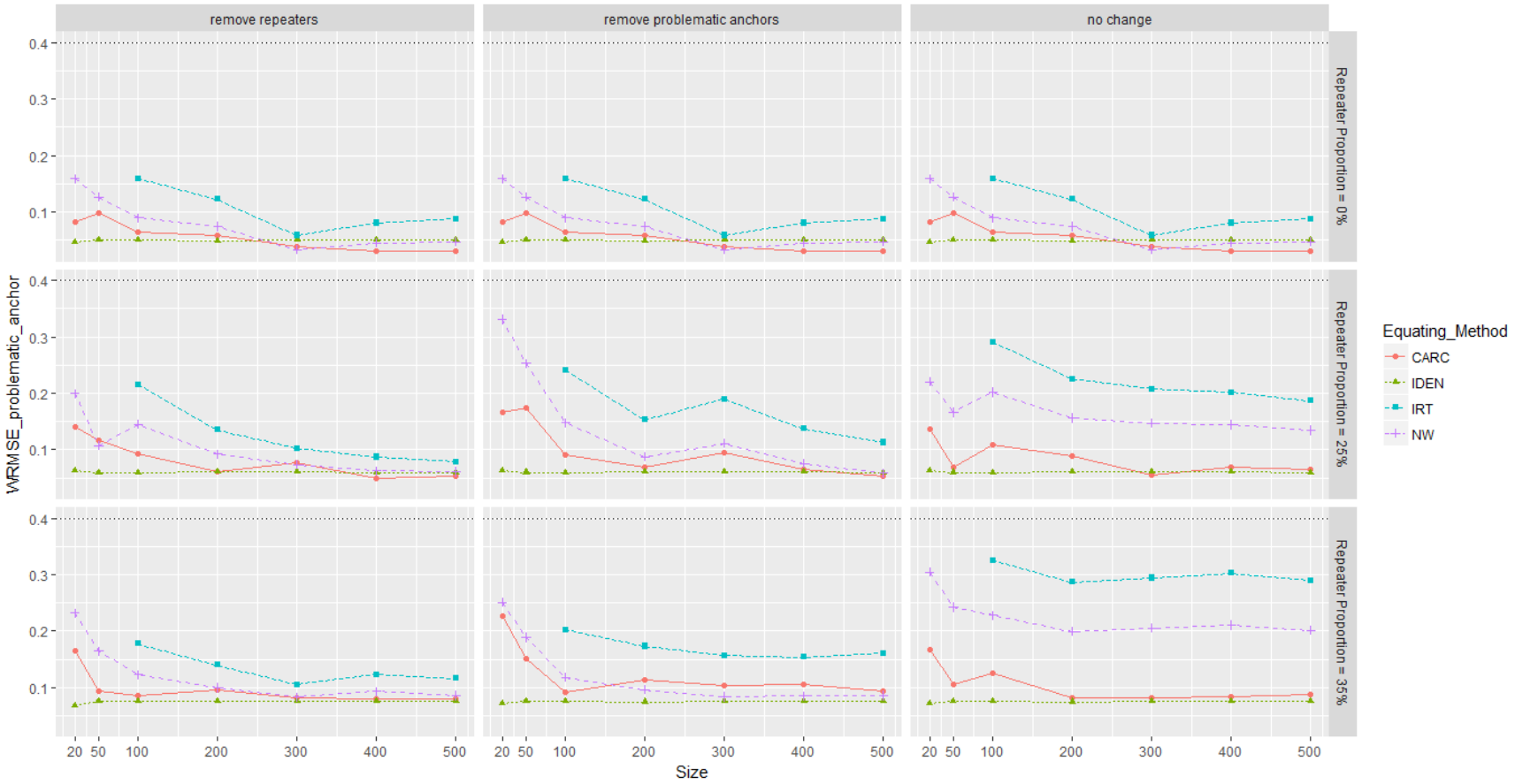
Figure 4.25. WRMSE of Non-problematic Anchor Test by Equating Methods

4.6.2 Problematic Anchor

Figure 4.26 shows values of WRMSE when equating with drifted anchor items. The line charts at the first column show the WRMSE under "removing repeaters" condition; the second column displays the line charts of "removing problematic anchor" condition and the charts at the third column portrays the WRMSE under "retaining repeaters" data management condition. The magnitude of WRMSE was generally low (smaller than 0.25) for the "removing repeater condition" with a mean of 0.09 (*SD* = 0.05). The means of WRMSE were 0.10 (*SD* = 0.07) and 0.12 (*SD* = 0.08) for "excluding problematic anchor" and "retaining all items and repeaters" conditions, respectively. The larger WRMSE at the third columns may result from the last two charts where 25% and 35% repeaters were included. For example, the mean value of WRMSE produced by Rasch equating was close to 0.35 when 35% repeaters were all remained before equating procedure. A closer visual inspection at each chart shows that circle-arc and identity equating stood out because of their robust and stable performance. The overall mean WRMSE produced by circle-arc equating, identity equating, Rasch equating and nominal weight mean equating were 0.08 (*SD* = 0.05), 0.06 (*SD* = 0.01), 0.15 (*SD* = 0.07) and 0.13 (*SD* = 0.08), respectively. This may indicate that Rasch equating and nominal weight mean equating produced higher overall RMSE than identity equating and circle-arc equating when the drift occurred to the anchor items. However, the performance of Rasch equating and nominal weight mean equating were not unsatisfactory across all conditions. The charts at first column indicate the WRMSE values provided by different equating techniques were closer at large sample size levels. In the second column, the difference among classical equating approaches was reduced as sample size increased.

Nominal weight mean even produced smaller WRMSE than circle-arc equating if $N >$ 300 and repeater proportion equal to 35%. Regarding the change across sample size levels, the mean of WRMSE across sample size levels range from 0.08 (SD = 0.05) at N = 500 sample size level to 0.16 (SD = 0.10) at N = 20. The WRMSE steadily decreased as sample size increased from 20 to 500.

The overall means for non-problematic and problematic anchor equating condition were 0.08 (SD = 0.05) and 0.10 (SD = 0.07), respectively. The charts also reflect that the problematic anchor test caused higher overall RMSE and larger variation. The Rasch equating and nominal weight mean equating were more sensitive to drift in anchor test and a large proportion of repeaters. The last two charts in the third column are the main sources leading to higher WRMSE under problematic anchor condition. Circle-arc equating and identity equating can produce more robust WRMSE across all test and equating conditions. If the presence of drifted anchor was unknown, perhaps the best solution to minimize the RMSE was removing all repeaters. Retaining all repeaters and drifted anchor items may yield the largest volume of WRMSE.

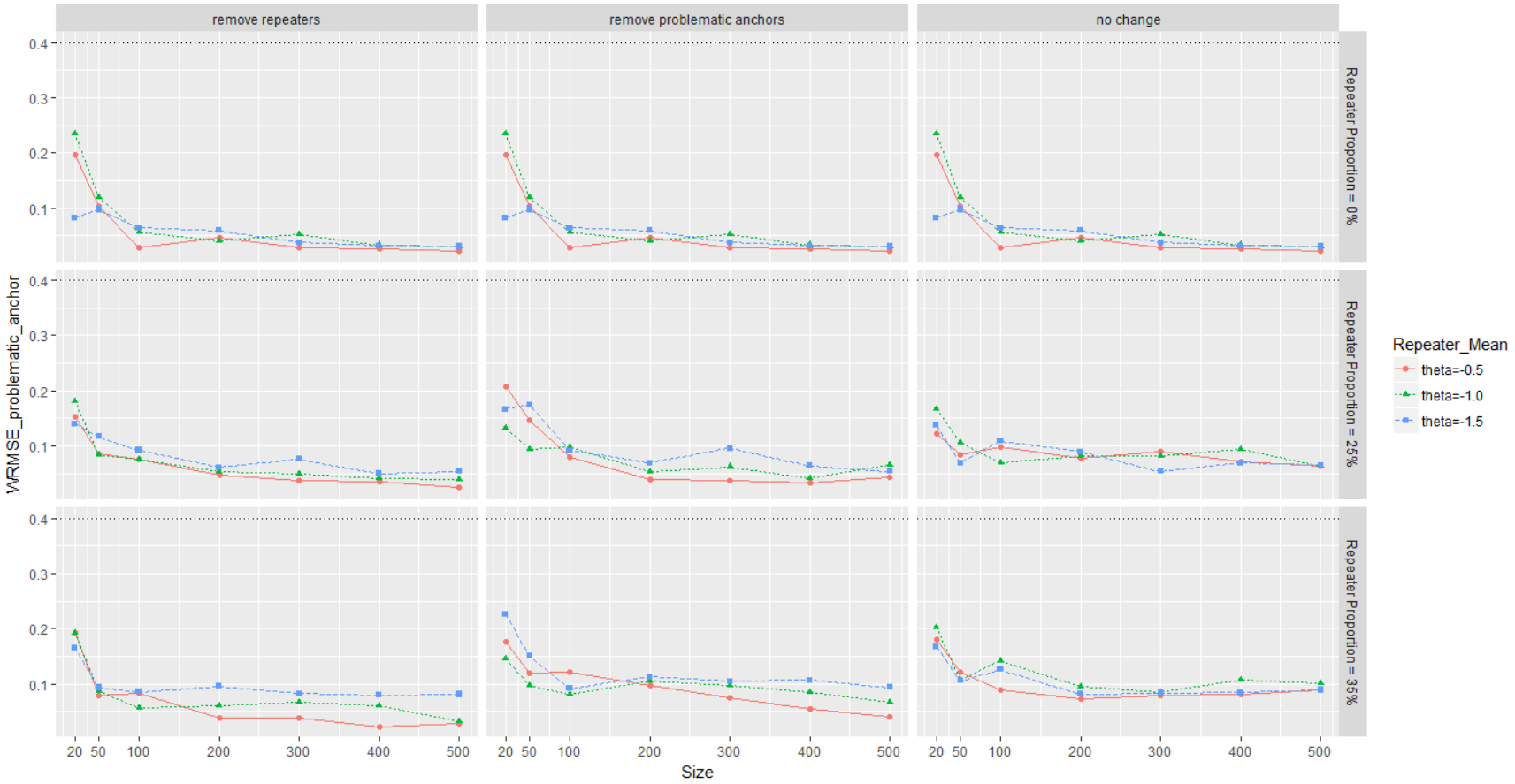


Note. Distribution of repeaters: $\theta_{R3} \sim N$ (-1.5, 1)

Figure 4.26. WRMSE of Problematic Anchor Test by Equating Methods

4.6.3 Repeater Mean

The mean and standard deviation of WRMSE among three repeater distributions show that repeater mean was a weak but non-negligible factor in influencing the WRMSE. The mean and standard deviation for $\theta_{R1} \sim N$ (-0.5, 1), $\theta_{R2} \sim N$ (-1.0, 1) and $\theta_{R3} \sim N$ (-1.5, 1) are 0.08 (SD = 0.05), 0.08 (SD = 0.05), 0.09 (SD = 0.05), respectively. Under drifted anchor condition, the mean and standard deviation for $\theta_{R1} \sim N$ (-0.5, 1), $\theta_{R2} \sim N$ (-1.0, 1) and $\theta_{R3} \sim N$ (-1.5, 1) were 0.09 (SD = 0.07), 0.11 (SD = 0.07), 0.11 (SD = 0.06), respectively. Figure 4.27 shows that removing 35% repeaters with mean of -0.5 can result in lower WRMSE values. Therefore, both summary statistics and line charts indicate that repeater mean closer to non-repeaters can provide lower WRMSE. However, the magnitude of reducing WRMSE was not remarkably large.



Note. Equating Method: Circle-Arc Equating

Figure 4.27. WRMSE of Problematic Anchor Test by Repeater Mean

4.7 Effect on CDC

Conditional difference curve (CDC) is an indication of population invariance that displays the difference between subgroup (non-repeater) and the total group equating function. Recall the formula to compute CDC in the previous chapter:

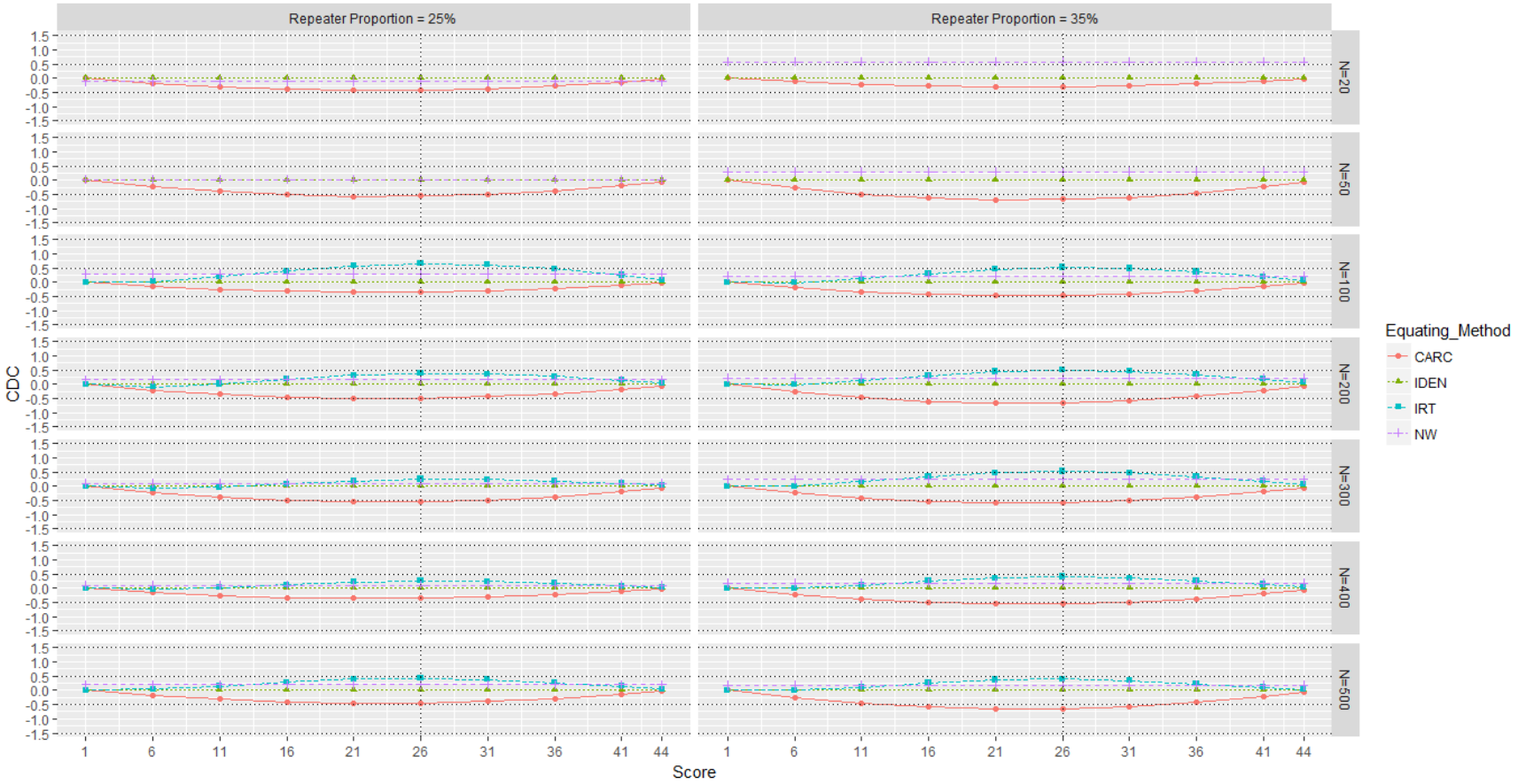
$$CDC(x) = e_{Pj}(x) - e_P(x). \quad (3.40)$$

If the value of CDC was negative, it means the equated score resulted from non-repeaters were lower than the total group; otherwise, the total group produced higher equated score than the non-repeater group. The magnitude of CDC was evaluated by the difference that matters (DTM) to determine the level of violation to invariance property. In the current study, DTM is equal to 0.5 which is a half unit of scaled score. If the value of CDC falls into the range between -0.5 to +0.5, the property of invariance might not be threatened. The graphs portray how CDC varies across score scale while Table B12 and Table B13 show the value of CDC at cut-score point (x = 26). One thing should be noticed is identity equating always provided a CDC equals to 0 because the equated score produce by identity equating did not differ between different examinee groups.

4.7.1 Non-problematic Anchor

Figure 4.28 displays the patterns of CDC across score points. The magnitude of CDC reaches to the highest at the cut-score point. Among three equating techniques, circle-arc equating tended to produce negative CDC, indicating non-repeater group had a lower equated score than the total group. In other words, retaining repeaters might make the equated score higher and therefore more examinees were likely to pass. The magnitude of CDC provided by nominal weight mean equating was slightly smaller than

circle-arc equating and close to the value of zero. Rasch equating was likely to produce positive CDC, indicating non-repeater group was likely to produce a higher equated score than total sample group. The zero lines are highlighted in green, which also represents the CDC values provided by identity equating. In Figure 4.28, the lines represent nominal weight mean equating and Rasch equating are very close and overlap at the cut-score point if $N \geq 200$. The patterns of three equating techniques were not very similar; however, the values of CDC remained in the DTM range under most of the test conditions. The mean CDC at cut-score point provided by circle-arc equating, Rasch equating and nominal weight equating were -0.25 (*SD*= 0.23), 0.19 (*SD*= 0.20), 0.12 (*SD*= 0.14), respectively. This confirms that circle-arc equating produced the largest negative CDC; Rasch equating provided second largest positive CDC and nominal weight mean equating provided the smallest positive CDC. The CDC values might be invariant of sample size but were influenced by the proportion of repeaters. Under 25% repeaters condition, the magnitude of CDC was within (-0.5, 0.5) range across all score points. Under 35% repeaters condition, the magnitude of CDC was slightly exceeding the DTM using circle-arc equating. The values of CDC did not differ across different sample size levels. However, at the sample size level of 20, the average CDC at the cut-point score was lower than other sample size levels. It does not mean the summary statistics are contradictory to our conclusion that CDC values are invariant to sample size. One possible explanation is that Rasch equating was not applied at $N = 20$ and $N = 50$ and this may impact the average CDC. The CDC values at sample size levels over 50 were stable across equating methods.



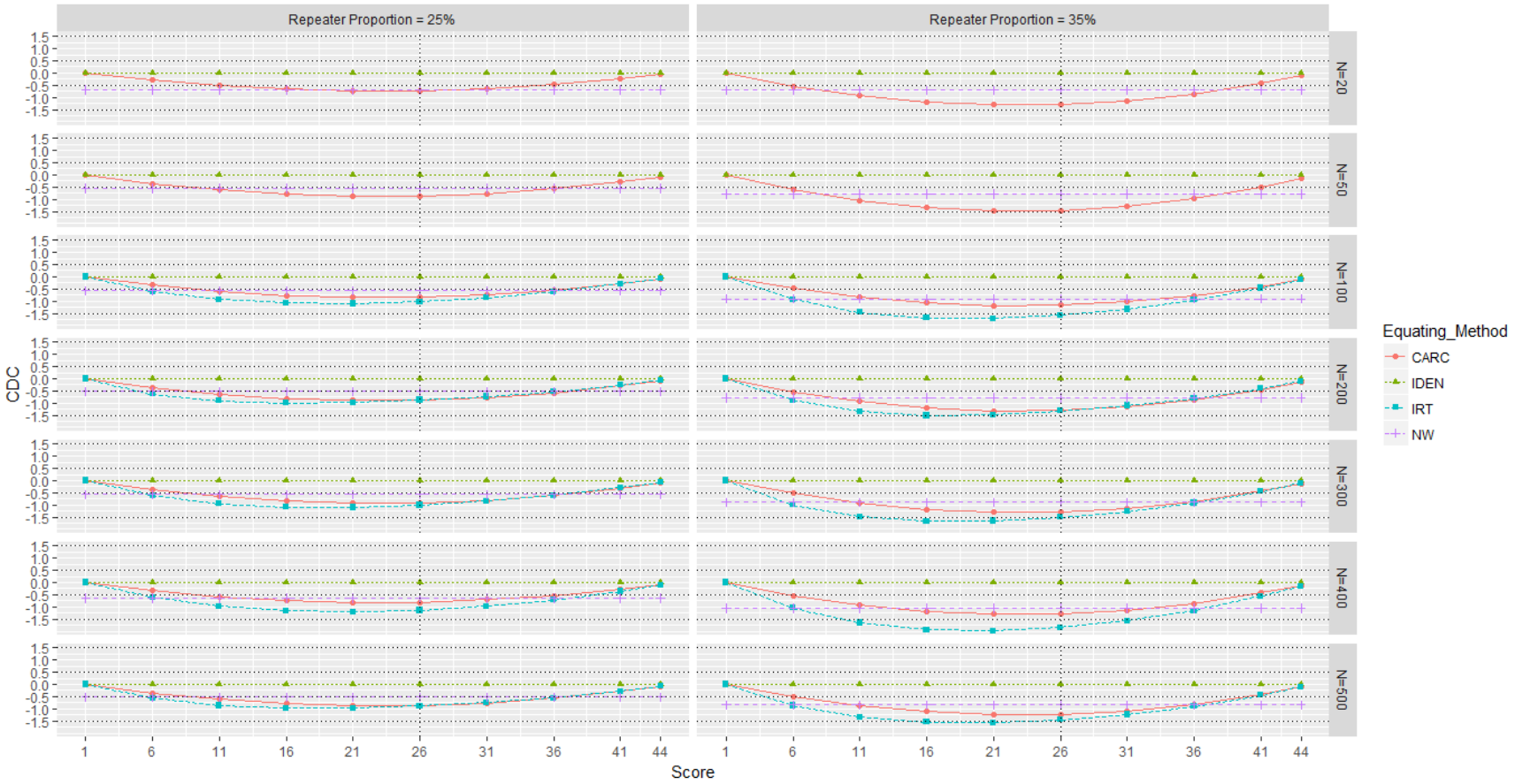
Note. Distribution of repeaters: $\theta_{R3} \sim N$ (-1.5, 1)

Figure 4.28. CDC of Non-problematic Anchor Test by Equating Methods

4.7.2 Problematic Anchor

Compare to the results from above subsection, the drift in anchor test caused a larger difference between total sample group and non-repeater group. In Figure 4.29, all techniques created negative CDC values under two repeater proportions. When sample size was larger than 100, Rasch equating provided the largest amount of CDC, especially around the middle of score scale. Both circle-arc equating and Rasch equating provided a U-shape curve wile nominal weight mean equating tended to produce a stable line across score points. Similar to the Figure 4.28, the sample size had little effect on CDC but the proportion of repeaters played a more important role. Increasing proportion of repeater can enlarge the magnitude of CDC values, especially at cut score point. The mean value of CDC at cut-score point for 25% and 35% repeaters are -0.82 (*SD* = 0.34) and -0.93 (*SD* = 0.34).

Several important findings can be drawn from Figure 4.28 and Figure 4.29. Firstly, increasing repeaters proportions could enlarge CDC. Equating with drifted anchor can lead to a large amount of CDC and therefore violating the invariance property of equating. Furthermore, repeater proportion had a direct effect on CDC and this effect was magnified under problematic anchor condition. Next, nominal weight mean equating provided the smallest magnitude of CDC values across test conditions. Lastly, the presence of drifted anchor was likely to produce negative CDC. That is, equated score resulted from the total sample was higher than that of the non-repeater sample, which consequently led to lower reported scores than total sample group. If the cut-score was fixed across all test administrations, equating with the non-repeaters group would lead to lower passing rate than total sample group.

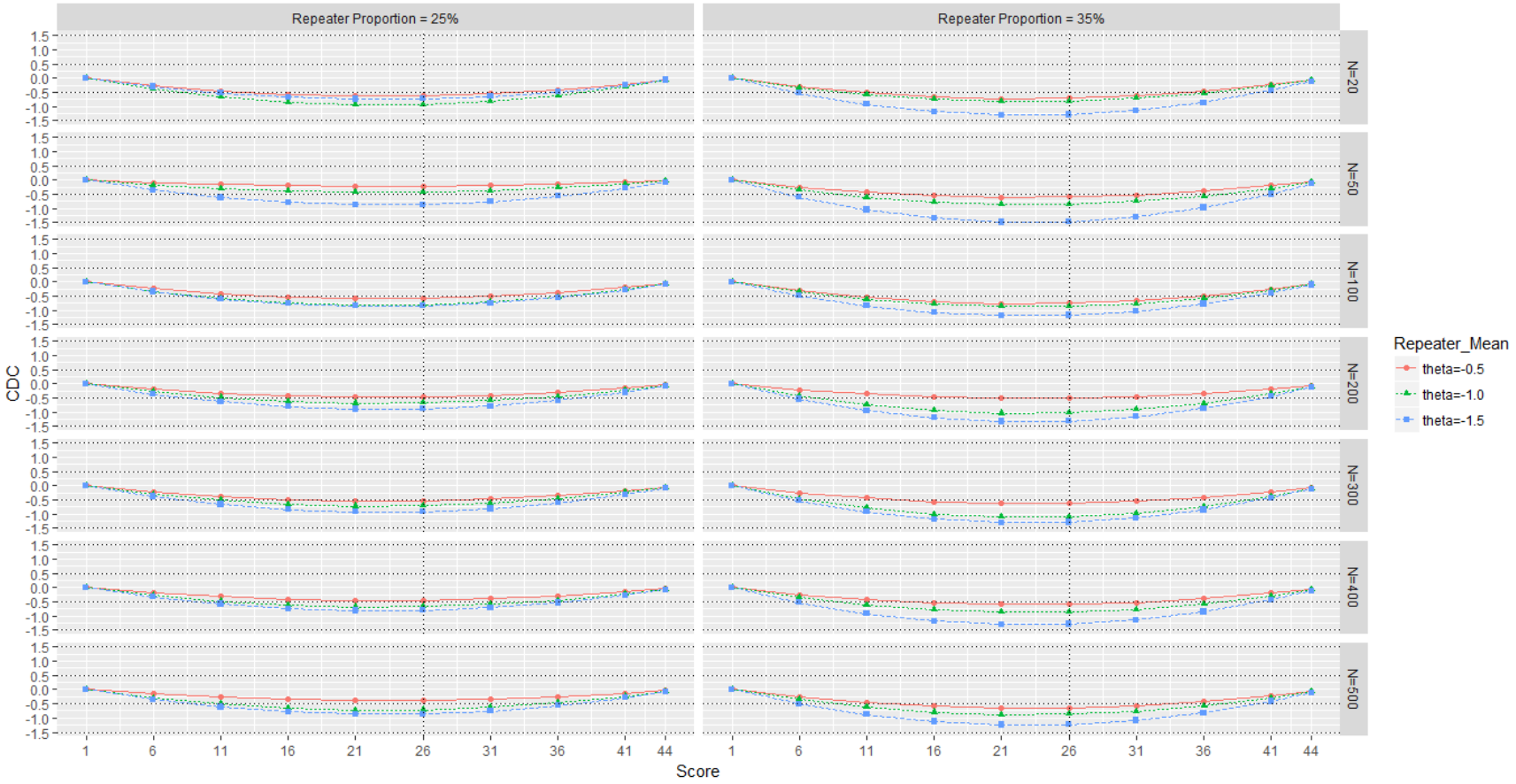


Note. Distribution of repeaters: $\theta_{R3} \sim N$ (-1.5, 1)

Figure 4.29. CDC of Problematic Anchor Test by Equating Methods

4.7.3 Repeater Mean

Figure 4.30 display the CDC values across repeater mean conditional on circle-arc equating techniques. Apparently, repeater mean had a strong impact on CDC values. The red line represents the repeaters follow a distribution of $\theta_{R1} \sim N$ (-0.5, 1) while blue line representing the distribution of $\theta_{R3} \sim N$ (-1.5, 1). Under distribution condition of $\theta_{R3} \sim N$ (-1.5, 1), the CDC values were more likely to exceed DTM. The summary statistics also confirm this result because the mean CDC at cut-score point decreased from -0.76 to -1.13 as repeater mean decreased from -0.5 to -1.5 under problematic anchor condition. Under non-problematic anchor condition, the mean CDC dropped from 0.00 to -0.15 as the mean of repeater decreased from -0.5 to -1.5



Note. Equating Method: Circle-Arc Equating

Figure 4.30. CDC of Problematic Anchor Test by Repeater Mean

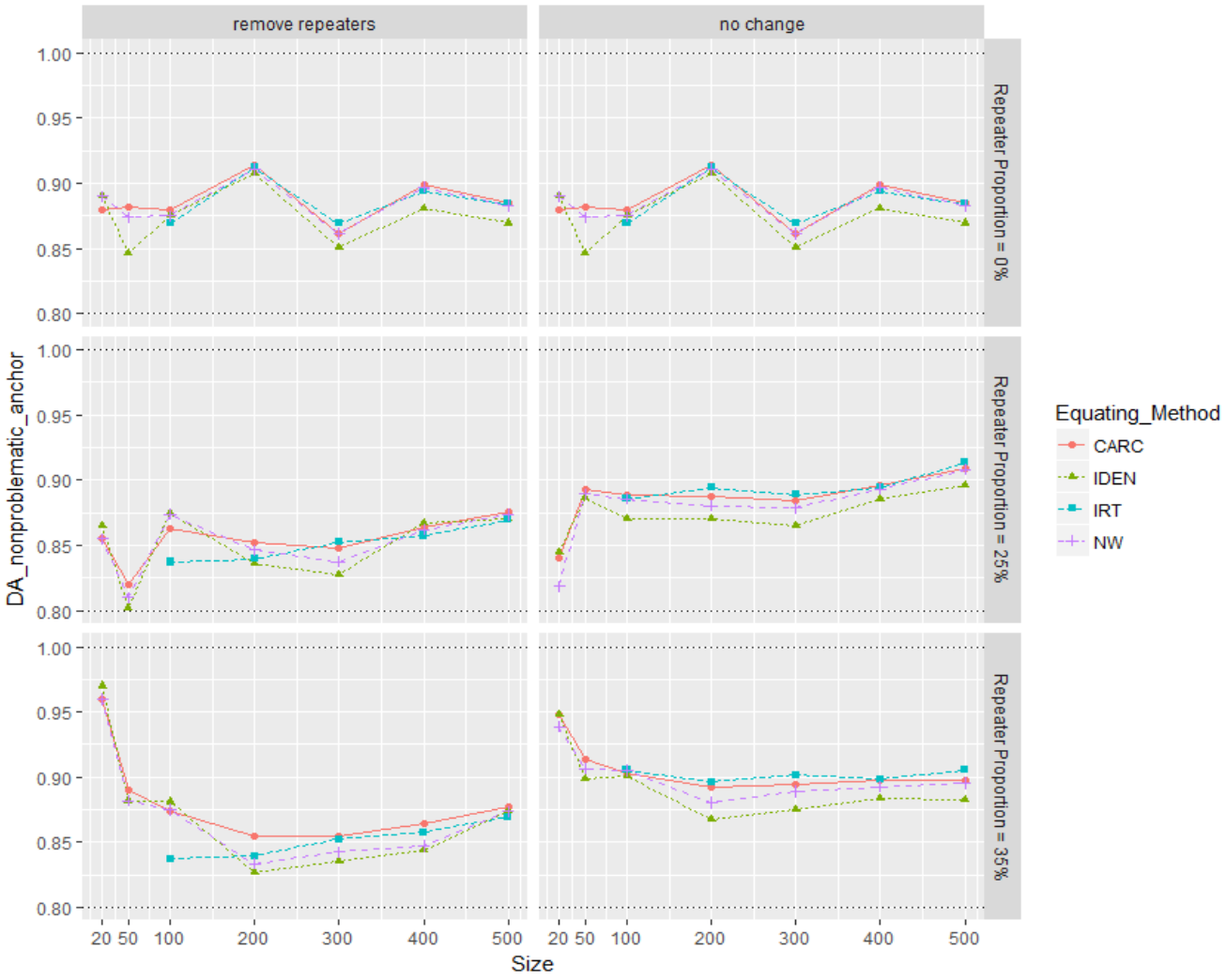
4.8 Effects on DA

Decision accuracy (DA) implies the agreement between true performance classification and estimated classification after equating. The DA is described in figures and tables. Figure 4.31 and Figure 4.32 explore how DA differed between equating approaches by fixing the repeater mean as $\theta_{R3} \sim N$ (-1.5, 1). Figure 4.33 is an example showing if there were differences between three repeater distributions in DA.

4.8.1 Non-problematic Anchor

Six charts are displayed in the Figure 4.31. The first column shows the DA values under "removing repeaters" condition and the charts at the second column show the DA values under "retaining repeaters" condition. Visual inspection implies that the DA resulted from original data set were slightly higher than the DA resulted from removing repeater solutions. The difference between two data management strategies was 0.01, which means 1% more examinees might be misclassified if repeaters were excluded. The better performance under "retaining repeaters" condition was consistent with WRMSB and WRMSE results. Holding data management strategy constant, the patterns of DA varied across repeater proportions. For 0% repeater proportion conditions, the DA values were bouncing between 0.85 to 0.90 across sample size levels. Under 25% repeater conditions, DA at N =20 or N =50 were bouncing between 0.80 to 0.85 and then suddenly arose when N = 100. The charts at the last row represent the condition where repeaters proportion was 35%. There was an elbow at the sample size level of 50 where the DA dropped from the highest values and then steadily decreases from 0.85 (removing repeaters) or 0.90 (retaining repeaters). The mean of DA across sample size levels ranged

from 0.87 (*SD* = 0.05) with *N* = 20 to 0.88 (*SD* = 0.02) at other sample size levels. The standard deviation of smallest sample level was almost double that of other sample size levels. The more variability in DA at the *N* = 20 sample size level in the Figure 4.31 also confirms with the summary statistics where variance was higher at lower sample size levels. Rasch equating, circle-arc and nominal method were likely to provide very close DA values, however, identity equating was likely to produce a slightly lower DA. The mean DA produced by circle-arc equating, identity equating, Rasch equating and nominal weight mean equating is 0.88 (*SD* = 0.04), 0.87 (*SD* = 0.02), 0.88 (*SD* = 0.02) and 0.88 (*SD* = 0.06).



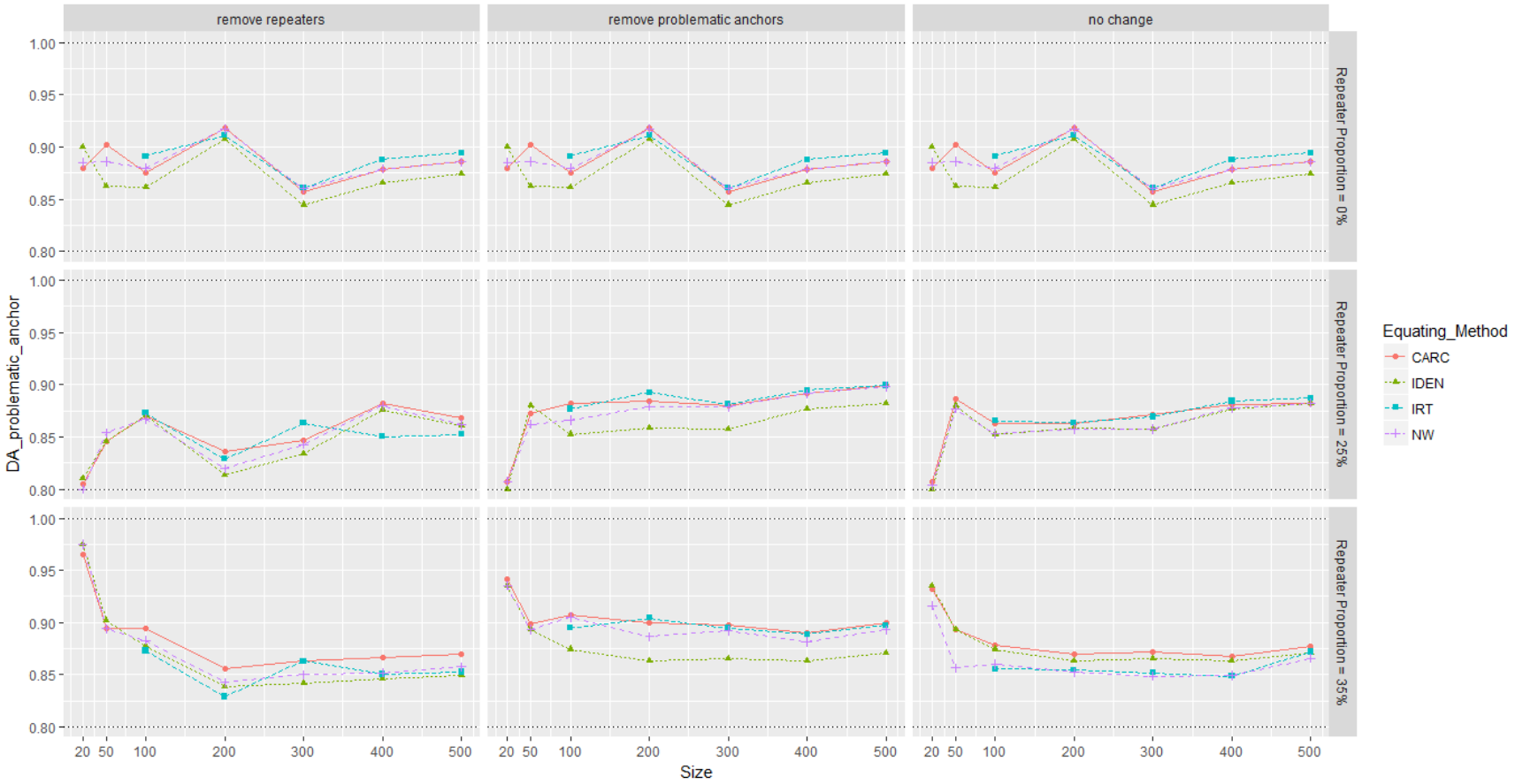
Note. Distribution of repeaters: $\theta_{R3} \sim N$ (-1.5, 1)

Figure 4.31. DA of Non-problematic Anchor Test by Equating Methods

4.8.2 Problematic Anchor

Figure 4.32 show values of DA when equating was performed with drifted anchor items. Similar to Figure 4.31, holding the data management strategy constant, 0% repeater proportion had a DA pattern that was bouncing around 0.85; 25% repeater proportion had a pattern that the DA was lower at sample size level of 20 or 50 and then stabilized at a value between 0.85 and 0.90; 35% repeater conditions resulted in a pattern that DA was the highest at smallest size, dropped to 0.90 at N =50 and then steadily decreased. Among three data management conditions, the "removing problematic anchor" condition had slightly higher DA than other conditions between the repeater proportions. The summary statistics also confirm this conclusion, that is, "removing problematic anchor condition" had 0.01 higher DA than other two conditions. A closer look at the DA at each chart reveals that identity equating tended to provide a slightly lower DA in most of the conditions. Circle-arc equating produced the highest overall DA across conditions (M = 0.88, SD = 0.02), which was 0.02 higher than the overall DA provided by identity equating.

In sum, the DA under two anchor tests conditions were similar in terms of the effects of repeater proportion, sample size, and equating techniques. Both figures had DA with different patterns across repeater proportion, large variation at smallest sample size level, and similar equating techniques performance.

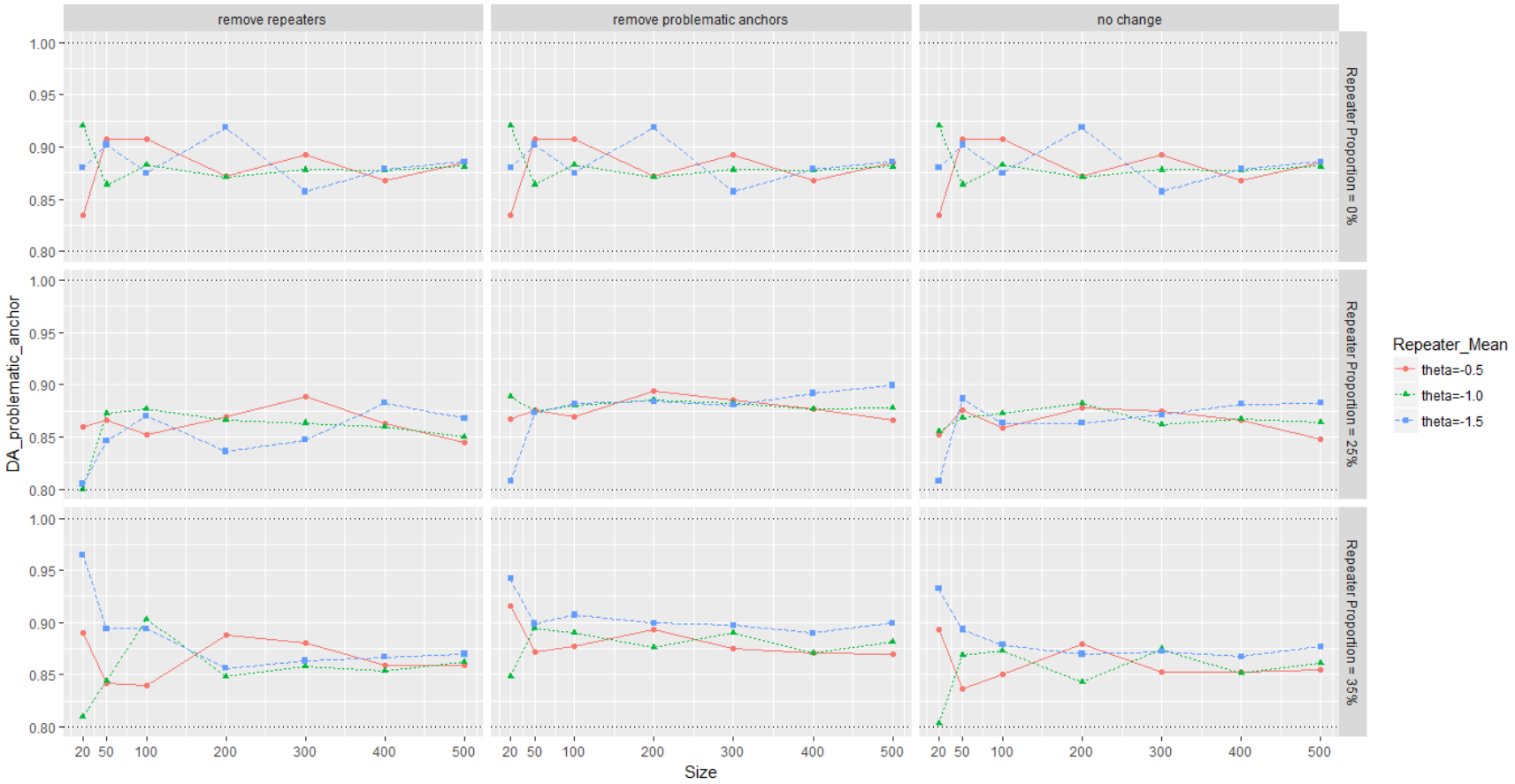


Note. Distribution of repeaters: $\theta_{R3} \sim N$ (-1.5, 1)

Figure 4.32. DA of Problematic Anchor Test by Equating Methods

4.8.3 Repeater Mean

The summary statistics in Table B14 and Table B15 do not reveal that the DA values were substantially different among repeater effect conditions. For example, under non-problematic anchor condition, the mean and standard deviation in for $\theta_{R1} \sim N$ (-0.5, 1), $\theta_{R2} \sim N$ (-1.0, 1) and $\theta_{R3} \sim N$ (-1.5, 1) were 0.88 (SD = 0.03), 0.87 (SD = 0.02), 0.88 (SD = 0.03), respectively. The Figure 4.33 shows the lines representing repeater means are entangling along the sample size levels, it is hard to conclude which line is always below or above the other two lines. The most striking feature in the figure is that the large variation was found at the N =20 sample size level. The DA values at this size level can reach above 0.95 or drop to 0.80. Because of the insufficient sample size, the raw score was less likely to be normally distributed. If more examinees were located at two ends of the score scale, the values of DA could be relatively high. If more examinees got scores at the middle of score scale that close to the cut-score point, the pass/fail decision might be less accurate and hence resulting in a low value of DA. This could be the reason why the variation was large at the smallest sample size level.



Note. Equating Method: Circle-Arc Equating

Figure 4.33. DA of Problematic Anchor Test by Repeater Mean

4.9 Final Note

The current study investigated eight evaluation criteria which reflect different aspects of equating results. Conditional bias, CSEE, conditional RMSE and CDC highlight the equating patterns across score scale whereas WRMSB, WSEE, WRMSE indicate the overall equating results. DA denotes the accuracy of the performance classification if the cut-score is 26. The conclusion that is drawn from these evaluation criteria is not always consistent. For instance, CDC results show nominal weight mean equating can most effectively retain the invariance property; DA results indicate circle-arc equating and Rasch equating outperformed other equating techniques; however, equating bias and errors imply that Rasch model performed less satisfactory than other equating techniques. This does not indicate the equating results based on different evaluation criteria were conflicting. Take WRMSB and DA as an example, the DA was computed based on individual's equated score while the computation of WRMSB was based on equating conversion table. The WRMSB indicates the overall bias of equating function along score scale, taking account examinee proportion at each score point and assuming each score point has an identical number of examinees. However, the data in this study were generated with a normal distribution where more examinees get scores in the middle scale than upper and lower ends. The pass/fail decision was made only based on one single score of each individual and compared this score to the cut-score. This decision made based on reported score cannot represent the equating accuracy of the entire conversion table that applies to all examinees across all score points. As a result, DA and WRMSB were associated somehow but differed in both computations and concepts. The final chapter would summary all equating results and discuss what

recommendations can be provided to testing programs with a small volume of examinees and large volume or repeaters. The recommendations would be made by synthesizing results derived from eight evaluation criteria.

CHAPTER 5

DISCUSSION

The current chapter consists of four sections. Firstly, the results of the study are summarized by each manipulating factor. Next, the first and second research questions are answered by discussing the performance of different small-sample equating techniques and the comparison of repeater effect solutions. The third section focuses on discussing the practical implication of equating bias and equating errors based on the accuracy of performance classification. The final section addresses the limitations of the current research and proposes some research ideas for further study.

5.1 Summary

This section provides the summary of equating results by each manipulating factor: sample size level, repeater proportion, repeater mean, drift in anchor test, repeater effect solutions and equating methods. Under each subsection, weighted average root mean bias (WRMSB), weighted average standard error of equating (WSEE), weighted average of RMSE (WRMSE), conditional difference curve (CDC) and decision accuracy (DA) are summarized. The interaction effects are also addressed if the outcomes were influenced by multiple factors.

5.1.1 Sample Size

The results show that equating bias was invariant to sample size and this is consistent with previous studies (Parshall et al., 1995; Sunnassee, 2011). However, the presence of larger variation at sample size levels of 20 and 50 may imply the magnitude of bias at small sample size levels can be more extreme than bias at larger sample size

levels. Similar results were found in DA, this may indicate that larger sample size conditions can provide more stable equating results. For standard error, standard errors decreased as sample size increased. Among three equating techniques, this pattern was more obvious in Rasch equating. Circle-arc equating was less sensitive to decreasing sample size but an "elbow" at sample size level of 100 can still be found in WSEE.

By combining the system error and random error of equating, size level of 200 was a "borderline" where the gap between equating techniques was more likely to remarkably minimize. In the other words, the disagreement between equating techniques was not significantly reducing as sample size level increased from 200 to 500. Therefore, size levels of 20, 50 and 100 were considered as small sample conditions that were distinguishable than the size of 200, 300, 400 and 500. In the following sections, the comparison between repeater effects solutions and equating techniques would be mainly discussed under size levels of 20, 50, and 100.

Summary statistics reveal that the overall mean values of DA was independent of size but the standard deviation s decreased as sample size increased. It is hard to conclude if the smaller sample size decreased DA because the value at N = 20 and N =50 levels can be either extremely high (DA = 0.95) or extremely low (DA = 0.80). The large variability indicates the small sample size can provide very unstable accuracy in pass/fail decision. For CDC, this measure was independent of sample size with respect to both mean and standard deviation. Furthermore, neither the CDC values at the cut-score point or patterns of CDC were influenced by increasing sample size, this feature was consistent across equating techniques.

5.1.2 Repeater Effects

The effects of repeater proportion and repeater mean were different depending on the evaluation measures. Among all evaluation criteria, repeater proportion and repeater mean had the least effect on the standard error of equating. Under the condition where the repeaters retained, there was a tendency that the WSEE was decreasing as the repeater proportion increased. One plausible explanation is that the sample size increased as more repeaters were included in the total equating sample. One piece of evidence to support this claim is that the WSEE remained constant across proportion levels if all repeaters were removed before equating.

Repeater proportion and repeater mean had a pronounced direct influence on CDC between the non-repeater group and the total sample group. The difference between two groups in equating functions enlarged as the repeater proportion increased and repeater mean deviated from zero. The results were not surprising because the distribution between the total group and the non-repeater group was getting more dissimilar as more repeaters included in the total sample group. The gap between groups can be amplified if the difference in mean proficiency levels between two groups grew further.

In terms of the influence to DA, repeater proportion was likely to impact the patterns of DA than the magnitude of DA. Under 0% repeater condition, DA was less deviated from the mean (DA = 0.88) across sample size levels compared to repeater proportion of 25% and 35%. When 25% and 35% repeaters were included in the equating procedure, the DA values were more extreme at N = 20 and N =50 levels and the most extreme values were found with repeater mean of -1.5 (see Figure 4.33).

The impact of repeater effect to equating bias was complicated because it was intervened with the problematic anchor, equating techniques and the approach to deal with repeaters. The most striking feature was found under the conditions where drifted anchor and repeaters were retained before equating. This feature was more obvious 35% repeater condition, in particular for Rasch equating and nominal weight mean equating. The bias produced by these two equating methods were distinctively higher than identity equating and circle-arc equating. However, under the conditions where repeaters or drifted anchor were excluded from the dataset, the influence of repeater proportion and repeater mean were attenuated.

5.1.3 Problematic Anchor Test

The drifted anchor resulted from item exposure to repeaters had a more prominent impact on CDC and equating bias than equating errors and classification accuracy. As it is discussed above, the equating errors were strongly influenced by sample size and DA had different patterns between repeater proportion conditions. Compare to CDC and bias, these two evaluation criteria were more invariant under non-problematic/problematic anchor conditions. One the contrast, CDC was likely to magnify under drifted anchor condition, which led to values exceeding DTM and threatening the property of equating invariance. Under problematic anchor condition, the level of violation to invariance property was different between equating techniques. The nominal weight mean equating was more likely to retain the invariance property whereas the Rasch equating tended to provide a high CDC values that were greater than DTM range. Conditional on same test conditions, including problematic anchor can consistently increase the equating bias across equating techniques. However, the degree of impact differed between equating

techniques. The performance of equating techniques would be compared at the last subsection.

5.1.4 Solutions to Mitigating Repeater Effects

Three solutions to mitigate repeater effects were proposed and examined in the current study. These solutions were removing repeaters, excluding drifted anchor that were exposed to repeaters and applying IRT equating to retain the invariance property. The first solution was applied to the data management strategy where the repeater responses were removed from the original the data set. The second solution was conducted when equating was performed with anchor test without drifted anchor items. The third solution was nested within the first two solutions. To highlight the effects of these solutions, equating was performed on original the data set where all responses and anchor were retained.

If the anchor test was not drifted, removing repeaters did not significantly improve equating accuracy of classification decision. In contrast, retaining repeater responses can slightly improve equating accuracy and DA. In addition, the main improvement was caused by circle-arc equating, nominal weight mean equating and Rasch equating. The identity equating remained the same regardless of the change to the data. If the anchor was drifted, removing repeaters provided a lower overall equating bias than excluding problematic anchor items.

The goal of implementing Rasch true score equating was to retain the property of invariance, which was reflected by the CDC resulted from the difference between the total group and the non-repeater group. However, Rasch TSE did not have an outstanding performance in reducing the difference between groups. Furthermore, Rasch equating

produced the highest biased CDC if the drifted anchor was involved in equating procedure. Therefore, applying Rasch equating was not an ideal solution to retain the invariance property.

Compared to the removing repeater solution, excluding problematic anchor can retain more information because all responses were kept in the data set. Excluding problematic anchor can result slightly lower overall equating bias and errors than excluding repeater responses. Regarding classification accuracy, excluding drifted anchor could result in overall higher DA values than other data management solutions.

In sum, the results of equating bias and errors suggested that removing repeaters or contaminated anchor test can improve equating results if the equating anchor was drifted due to repeaters. However, the performance of different equating techniques was not consistent across data management conditions and hence the performance of repeater effect solution depended on the specific equating technique. For example, circle-arc equating could provide a stable equating accuracy regardless of the changes made to the original data whereas nominal weight equating should be performed after the contaminated responses were removed. The comparison between small-sample equating techniques was described in the following section.

5.1.5 Equating Methods

The small sample equating techniques used in the current study were: circle-arc equating, nominal weight mean equating, identity equating and Rasch equating. Rasch equating was conducted when the sample size was equal to or larger than 100 because IRT modeling has a high requirement for the mixed-format test. At the sample size level of 100, Rasch equating produced distinctive higher SEE and RMSE. However, Rasch

equating was likely to produce a similar amount of error as the sample size increased from 300 to 500. Nominal weight mean equating provided lower overall SEE than Rasch equating. One thing should be noticed is that the nominal weight mean equating had higher conditional SEE than Rasch equating at upper and lower score scale. Circle-arc equating provided lower standard errors under most of the test conditions.

The comparison in equating bias and RMSE was more complicated because equating methods had different performance across test conditions. If there are no repeaters in the total sample group, all equating techniques provide a similar amount of bias yet identity equating had slightly higher bias than other equating methods. If there were repeaters but no drifted anchor, the amount of bias resulted from different equating techniques was close to each other. If the anchor was drifted because of the item exposure to repeaters, Rasch equating provides the highest bias among all equating techniques. The nominal weight mean equating provided less biased equating results than Rasch equating but was still distinctive higher than other equating techniques. Circle-arc equating had higher bias than identity equating when the sample size was smaller than 50; however, at large sample size levels, circle-arc equating can give the least biased results among all equating methods. Because of the influence of standard error, the RMSE shows that Rasch equating produced higher RMSE than other equating techniques. When the repeaters and contaminated anchor test were included in the equating, Rasch equating and nominal weight mean equating provided significantly higher RMSE than circle-arc equating and identity equating. For most of the conditions, especially the sample size levels of 20 and 50, the sequence of equating technique that was ordered by decreasing RMSE is Rasch equating, nominal weight mean, circle-arc equating and identity

equating. If the sample size was larger than 50, circle-arc equating produced similar and even smaller RMSE than identity equating. One thing should be noticed is that RMSE is the combination of SEE and bias and identity equating always provides zero standard error. If bias is considered as the only indication of equating accuracy, nominal weight mean and circle-arc provide a similar amount of equating bias if equating was performed with no drifted anchor. By synthesizing the equating bias and errors, circle-arc equating and identity equating had the most stable performance across all test conditions. Circle-arc equating slightly outperformed at size levels larger than 50.

The initial purpose of applying Rasch equating was to retain the property of invariance so that equating function would not differ between the non-repeater group and total sample group. Unfortunately, IRT did not give an outstanding performance under non-problematic anchor condition and even produced the highest level CDC if anchor drifted. Among three equating techniques, the nominal weight mean equating had the most satisfying results that may reduce the violation to invariance property.

The performance of equating methods on classification accuracy was different from the evolution criteria measuring equating accuracy or group difference. It is interesting that Rasch equating, circle-arc equating and nominal weight mean equating had a similar level of classification accuracy under most sample size levels whereas identity equating had overall lower DA. The most notable difference between identity and other equating was found under excluding problematic anchor test solutions.

5.2 Conclusion

This section provides the conclusion of the study by answering first two research questions. The first research question focuses on the comparison between three repeater

effect solutions by holding other testing conditions same. Additionally, the conclusion to the first question would answer the problem of violation invariance property. The second subsection answers the question regarding the comparison among equating techniques. This subsection also discusses if the equating results from equating techniques differed under varied test conditions.

5.2.1 Research Question 1

Research Question 1: How do different repeater effects solutions impact the equating results, holding other conditions constant? Does the exclusion of repeater approach hold the invariance property? Which solution(s) can produce higher equating accuracy and lower equating bias?

In answering the first research question, different repeater effect solutions do produce different level of equating bias and errors. In addition, the conclusions that are drawn from equating results may change depending on the existence of problematic anchor items. If repeaters were the only concern and the anchor test was not contaminated by drift items, retaining repeater or excluding repeater may not cause substantially different accuracy level. However, it is hard to guarantee that no drift occurs to items. If repeaters memorized some anchor items and retook the same set of anchor items in the new form test, these exposed items might appear easier to repeaters than other non-repeaters. Under this circumstance, it is always better to take actions than insisting performing equating with drifted anchor items and all repeater response.

By removing repeater's responses, all drifted anchor caused by repeaters were precluded at the same time. The effects of repeaters and problematic anchor were both eliminated at the same time. However, the main limitations of this solution were reducing

sample size and violation to population invariance property. As it is discussed, excluding 25% - 35% examinees from the total sample can significantly reduce a large of examinees, especially to small volume testing programs. Additionally, the difference was amplified as the drifted anchor involved in the equating procedure. Although nominal weight mean equating can slightly reduce the threats to property invariance, the equated score derived from repeater group was still likely to be lower than the equated score of the total group, which may lead a lower passing rate than equating with total sample group.

If most of the drifted anchor can be detected and eliminated, all examinees can retain at the total group and the invariance property can be held in terms of repeaters and non-repeaters. In addition, this solution can give higher classification accuracy. Unfortunately, detecting drifted anchor is more difficult than detecting repeaters in reality. The problems can become more complicated if the drift occurs on both anchor and non-anchor items. It is difficult to completely remove all drifted anchor that only due to exposure. Thus, removing repeaters is more straightforward and simplified than detecting problematic anchor test.

In sum, retaining all repeaters is suggested if no anchors are drifted. If the exposed anchor items caused the severely drift, removing repeaters can lead to less equating bias and errors despite that excluding problematic anchors can hold the invariance property regarding repeater groups and provide slightly higher decision accuracy level.

5.2.2 Research Question 2

Research Question 2: How do different small sample equating techniques impact the equating results, holding other conditions constant? Does the performance of equating techniques differ depends on test conditions and repeater effects solutions? If there are interactions, which test and equating conditions produce less equating errors and bias?

The performance of equating techniques differed depending on test conditions and data management strategies. There was no best or worst equating method under all test conditions. However, circle-arc equating had the most stable performance that was relatively invariant to sample size, anchor and repeater effects. At sample size level of 20 and 50, circle-arc equating and identity equating were able to provide stable equating results regardless of the number of repeaters and problematic anchor. The performance of Rasch equating and nominal weight mean equating might interact with data management conditions, problematic anchor and sample size. At larger sample size level ($N > 100$), Rasch equating and nominal weight mean equating can also provide more satisfactory equating results. The equating bias produced by all equating techniques were similar under most of the test conditions if repeaters or problematic anchors were removed. However, Rasch equating and nominal weight mean equating techniques were likely to produce higher standard errors than circle-arc equating across all sample size levels. In terms of the ability to hold invariance property, nominal weight mean equating produced the least CDC and attenuate the violation to invariance property at the cut-score point. Therefore, it is hard to conclude which equating method has the best performance across all conditions. For instance, circle-arc equating had the most impressive performance

under most of the conditions, identity equating was likely to provide the least bias at smallest size level.

Although Rasch equating result in high equating error among all equating techniques, it can produce less WRMSB than identity if no repeaters are included at size levels larger than 50. When N = 500 with 0% repeaters, there was a trivial difference in bias between circle-arc equating, Rasch equating and nominal weight mean equating. These three techniques can produce lower bias than identity equating. The comparison between equating techniques would be made at given conditions at following sections.

5.3 Practical Implication and Recommendation

The practical implication of the study is discussed by answering the third research question. The practical implication is revealed by one important consequence: the accuracy of classification based on the reported score. Thus, the section 5.3.1 would report the implication mainly based on the classification rates between true and estimated classification and how to associate the DA with other equating measures. The second subsection discusses the recommendation regarding equating methods, repeater effects under conditions with given sample size and proportion of repeaters.

5.3.1 Practical implication

Research Question 3: What are the practical implications of this study? How do the equating results and population invariance affect performance classification at the individual level? At a given condition of sample size and proportion of repeaters, what recommendations should be given to get an acceptable level of results?

Classification accuracy is one of the reliability index measuring the degree of accuracy of the performance classification made based on the reported score. Unlike equating bias and equating errors, DA can directly reflect the percentage of examinees that are misclassified or correctly classified at the individual level. Reporting classification/misclassification rate can reveal the consequences regarding decision making for individuals. As a result, the practical implication of the current study is discussed by reporting the misclassification rates across small sample equating techniques and repeater effect solutions.

If the test had no repeaters and no drifted anchors, the total errors (random and systematic errors) could lead 7% - 15% misclassification. That is to say, even under the ideal conditions with sufficient sample size. Performing equating can still produce inevitable errors and cause certain amount misclassifications. Although it is hard to totally eliminate misclassification, it is possible to minimize the errors and increase the classification accuracy. If anchor test was not drifted, retaining all examinees can improve 5% classification than removing repeaters when the sample size was larger than 50. The misclassification rate would not change if the repeater proportion increase from 25% to 35%.

Under the condition with drifted anchor, removing problematic anchor using nominal weight mean, circle-arc and Rasch equating could keep the misclassification rate around 13% for most of the sample size levels. However, if the drifted anchor cannot be completely eliminated, retaining all responses and problematic anchor would result in a misclassification rate around 15% for sample size over 20. Under the same test conditions, removing all repeater responses can keep the misclassification within an

interval between 10% to 17% for sample size larger than 20. The variation was large at the smallest sample size level, misclassification rate could reach to 20% with 25% repeaters and then dropped to 5% with 35% repeaters. The large variability in misclassification between repeater proportions indicated that the overall classification rate could be strongly influenced by few extremely low or high DA values. Under $N = 20$ level, only 4 repeaters were added from 25% to 35% conditions. It is very likely that the classification accuracy was very unstable and sensitive to the changes of the total sample.

In general, the results of DA were associated but not always agreed with equating bias and errors that are reported in the previous sections. The DA gives a better understanding of how the overall errors impact outcome at the individual level. The following section would make some suggestions given certain test conditions based on the results regarding equating accuracy and practical implication with respect to DA.

5.3.2 Recommendation

In the current study, the highest proportion repeater level was 35% in the new test form. In other words, if a small volume testing program only gets 20 examines in one test administration, almost 13 repeaters could be removed from the total group using the first repeater effect solutions. Under this circumstance, removing examinees or applying equating model requiring large sample size is not recommended. In other words, the best solution is to keep as much information as possible and apply the equating methods that required small sample size. In terms of the recommendation for repeater effects, it is suggested to keep repeater responses and exclude drifted anchor items if they can be detected. For equating methods, circle-arc equating and identity are suggested because they can provide remarkably lower bias and errors at $N = 20$ and $N =50$ sample size

levels. If the sample size is smaller or equal to 50, identity equating is recommended because applying identity equating at smallest sample size level can reduce the standard errors and retain the equating invariance.

At the size level of N =100, which is a borderline that determines if the IRT equating can be applied, the repeater effects had a weaker influence on equating accuracy than the choice of equating methods. Since the N = 100 is sufficient for most of the equating methods, removing all repeater responses can avoid the misuse of the problematic anchor. However, classical equating is still recommended over IRT equating with 100 examinees participating the test. Nominal weight mean equating and circle-arc can produce satisfactory results yet circle-arc equating is slightly preferred with a smaller amount of errors and bias.

As long as the sample size is larger than 100, removing repeater is still suggested. The difference in evaluation criteria between equating methods is getting smaller. If the testing program strongly demands the properties of IRT modeling such as population and item independence, Rasch equating might be applicable with a sample size over 100 yet highly recommended under conditions with at least 400 examinees (Kolen and Brennan, 2004). Although the equating evaluation criteria in this study indicate Rasch equating had a satisfactory performance, it did not imply that sample size of 100 or 200 was sufficient for mix-format tests. The accuracy and bias of parameters estimated by Rasch modeling and PCM were not examined in this study. As a result, IRT equating is not recommended for the test with small sample size less than 400.

5.4 Limitations

This study has some limitations regarding simulation design, data management, and selection of equating techniques.

Firstly, the performance of repeater is not easy to simulate. The current study simulated a scenario that repeaters had lower academic proficiency level than non-repeater groups. However, it is not uncommon that repeaters made some progress after their attempt and had a similar or even higher proficiency level. The results and conclusion of this study might change if repeaters had higher proficiency levels. Furthermore, the current study assumed the repeaters and non-repeater ability only differed in the mean but had same distribution shape. However, the reality is more complicated because the distribution of repeaters might differ from the total group and non-repeater group in skewness, kurtosis and so forth. Finally, it is possible that repeaters, non-repeaters or total group are not normally distributed. The current study simulated the examinee's response based on the examinee ability (θ) that followed a normal distribution. All conclusions drawn from the current study might not fully apply to other test conditions if examinees ability (θ) is not normally distributed.

Secondly, the current study simulated an item exposure scenario; however, drift can be caused by other reasons such as student growth, group difference, or interactions between examinees and test. Even the current study assumed that item exposure was the only reason caused the drift, the drift could be associated with the changes in item discrimination parameters as well as item guessing parameters. Lastly, the current study only considered a drift with $b_1(V) - b_2(V) = 0.50$. Different levels of drift need to be

considered in the further study to fully examine how drifted anchor intervened with repeater effects.

The third limitation is related to the second limitation. The current study fixed the problematic anchor that were exposed and memorized by repeaters and assumed that these items can be detected and totally removed. Although the exposure items can be detected in the application, removing these items from equating is not the most optimal solution. Anchor test should be representative of the total test, removing items of the anchor test might cause the anchor test less representative and lead to equated score with high bias. What is more, it is hard to know which are the true drifted anchor items that were only caused by exposure. The anchor items might be drifted because of the change of examinee group. Therefore, the solution of removing problematic anchor is not very realistic in the application unless there is strong evidence that some of the anchor items were drifted due to exposure.

The following limitations are associated with the equating design of the study. The current research only examined three classical small sample equating techniques and one IRT equating method that does not require a large number of examinees. However, these newly developed equating techniques are not the only equating techniques for the small-volume sample. Linear equating and pre-/post smoothing techniques for equipercentile equating also had satisfactory performance (Livingston, 1993; Han, Zhang & Colton; 1994, Parshall et al., 1995). It is possible these traditional equating techniques outperform the small-sample equating techniques that were investigated in the current study.

Moreover, the equating procedures were performed between two parallel forms under NEAT design. However, equating might be performed over multiple forms if testing programs administrate the test frequently. A sequence of equating might involve multiple forms built with similar content and statistical specifications where the cumulative equating errors should be computed across forms (Guo, 2010). Under this circumstance, the performance of equating techniques might be different from the condition where only two parallel forms were included.

Lastly, the criterion equating results were derived from equipercentile equating based on 5000 examinees. The equipercentile equating is always considered as a "gold standard" in previous studies (e.g., Livingston, 1993; Skaggas, 1995; Kim & Livingston, 2010; Albano, 2015). However, equipercentile equating can produce unavoidable equating bias and errors. As a nonlinear equating method, it might produce similar results as circle-arc equating (Livingston & Kim, 2008). In another word, there is a possibility that circle-arc equating had a satisfactory performance due to the similarity to the criterion equating method. If the true equating results were derived from a linear equating method, the conclusion might change.

As a result, further study may focus on the response data that are randomly drawn from the empirical data set. By doing so, distributions of repeater and non-repeaters would be more realistic, and the conclusions drawn from empirical data would be more representative than simulated response. Furthermore, it is important to compare the newly developed equating with more traditional equating methods such as linear equating or smooth techniques. Finally, it would be interesting to examine the equating results obtained from a sequence of equating across multiple forms.

APPENDIX A ITEM PARAMETERS

Table A1. Item Parameters of MC items (no problematic anchor items)

	Old Form (Y)			New Form (X)			Anchor Form (V)		
Item	a_1	b_1	c_1	a_2	b_2	c_2	a	b	c
1	1.39	-0.07	0.19	1.39	-0.07	0.19	1.39	-0.07	0.19
2	0.98	-0.74	0.19	0.98	-0.74	0.19	0.98	-0.74	0.19
3	1.27	-0.03	0.14	1.27	-0.03	0.14	1.27	-0.03	0.14
4	0.53	-0.14	0.21	0.53	-0.14	0.21	0.53	-0.14	0.21
5	1.02	-0.06	0.17	1.02	-0.06	0.17	1.02	-0.06	0.17
6	0.54	0.03	0.15	0.54	0.03	0.15	0.54	0.03	0.15
7	0.99	-0.31	0.06	0.99	-0.31	0.06	0.99	-0.31	0.06
8	1.11	0.41	0.03	1.11	0.41	0.03	1.11	0.41	0.03
9	1.57	-0.19	0.18	1.57	-0.19	0.18	1.57	-0.19	0.18
10	2.03	-0.01	0.28	2.03	-0.01	0.28	2.03	-0.01	0.28
11	1.67	-0.19	0.18	1.67	-0.19	0.18	1.67	-0.19	0.18
12	0.91	-0.72	0.16	0.91	-0.72	0.16	0.91	-0.72	0.16
13	0.58	-0.29	0.10	1.09	-0.44	0.30			
14	1.54	-0.11	0.18	0.54	0.01	0.17			
15	0.84	0.06	0.29	2.07	0.09	0.28			
16	1.07	-0.54	0.30	1.00	-1.16	0.20			
17	0.55	0.01	0.17	0.99	-0.01	0.06			
18	2.03	-0.01	0.28	1.02	-0.06	0.17			
19	0.98	0.14	0.23	0.56	0.01	0.17			
20	0.98	-0.74	0.19	0.77	-0.04	0.29			
21	1.02	-0.06	0.17	1.99	-0.01	0.28			
22	0.73	-0.62	0.16	2.03	0.11	0.18			
23	1.65	-0.19	0.18	0.82	0.71	0.25			
24	2.03	-0.01	0.28	0.80	-0.62	0.16			
25	0.82	-0.21	0.06	1.66	0.41	0.20			
26	0.76	0.07	0.20	1.24	-0.54	0.30			
27	1.67	-0.19	0.18	1.44	-0.37	0.33			
28	1.44	0.51	0.19	0.71	0.07	0.20			
29	0.73	-0.62	0.16	1.18	-1.16	0.18			
30	1.54	0.41	0.21	1.57	-0.15	0.18			
31	0.79	0.71	0.25	0.79	0.62	0.16			
32	1.01	-1.16	0.18	1.00	-0.62	0.16			
33	1.41	-0.47	0.33	2.07	-0.74	0.20			
34	1.45	-0.89	0.16	1.46	0.41	0.21			
35	1.54	0.41	0.21	1.62	-0.37	0.33			
36	1.53	0.51	0.11	1.26	0.09	0.13			

Table A2. Item Parameters of MC items (6 problematic anchor items)

	Old Form (Y)			New Form (X)		
Item	a_1	b_1	c_1	a_2	b_2	c_2
1	1.39	-0.07	0.19	1.39	-0.07	0.19
2	0.98	-0.74	0.19	0.98	-0.74	0.19
3	1.27	-0.03	0.14	1.27	-0.03	0.14
4	0.53	-0.14	0.21	0.53	-0.14	0.21
5	1.02	-0.06	0.17	1.02	-0.06	0.17
6	0.54	0.03	0.15	0.54	0.03	0.15
7	0.99	-0.31	0.06	*0.99*	*-1.31*	*0.06*
8	1.11	0.41	0.03	*1.11*	*-0.59*	*0.03*
9	1.57	-0.19	0.18	*1.57*	*-1.19*	*0.18*
10	2.03	-0.01	0.28	*2.03*	*-1.01*	*0.28*
11	1.67	-0.19	0.18	*1.67*	*-1.19*	*0.18*
12	0.91	-0.72	0.16	*0.91*	*-1.72*	*0.16*
13	0.58	-0.29	0.10	1.09	-0.44	0.30
14	1.54	-0.11	0.18	0.54	0.01	0.17
15	0.84	0.06	0.29	2.07	0.09	0.28
16	1.07	-0.54	0.30	1.00	-1.16	0.20
17	0.55	0.01	0.17	0.99	-0.01	0.06
18	2.03	-0.01	0.28	1.02	-0.06	0.17
19	0.98	0.14	0.23	0.56	0.01	0.17
20	0.98	-0.74	0.19	0.77	-0.04	0.29
21	1.02	-0.06	0.17	1.99	-0.01	0.28
22	0.73	-0.62	0.16	2.03	0.11	0.18
23	1.65	-0.19	0.18	0.82	0.71	0.25
24	2.03	-0.01	0.28	0.80	-0.62	0.16
25	0.82	-0.21	0.06	1.66	0.41	0.20
26	0.76	0.07	0.20	1.24	-0.54	0.30
27	1.67	-0.19	0.18	1.44	-0.37	0.33
28	1.44	0.51	0.19	0.71	0.07	0.20
29	0.73	-0.62	0.16	1.18	-1.16	0.18
30	1.54	0.41	0.21	1.57	-0.15	0.18
31	0.79	0.71	0.25	0.79	0.62	0.16
32	1.01	-1.16	0.18	1.00	-0.62	0.16
33	1.41	-0.47	0.33	2.07	-0.74	0.20
34	1.45	-0.89	0.16	1.46	0.41	0.21
35	1.54	0.41	0.21	1.62	-0.37	0.33
36	1.53	0.51	0.11	1.26	0.09	0.13

Table A3. Item Parameters of CR items

	Old Form(Y)			New Form(X)		
	a_1	b_{11}	b_{12}	a_2	b_{21}	b_{22}
1	0.70	-1.11	1.11	0.70	-1.11	1.11
2	0.88	-1.19	1.19	0.88	-1.19	1.19
3	0.60	-1.34	1.34	0.60	-1.34	1.34
4	0.36	-1.31	1.31	0.36	-1.31	1.31
Mean	0.63	-1.24	1.24	0.63	-1.24	1.24
SD	0.19	0.09	0.09	0.19	0.09	0.09
Min	0.36	-1.34	1.11	0.36	-1.34	1.11
Max	0.88	-1.11	1.34	0.88	-1.11	1.34

APPENDIX B. SUMMARY STATISTICS

Table B1. Summary Statistics for Reference Form Number Correct Score

Non-repeater	Non-repeater	Repeater Group1	Repeater Group2	Repeater Group3
Mean	26.91	23.30	20.16	17.07
SD	7.85	7.81	7.36	6.46
Median	27.00	23.00	19.00	16.00
Min	3.00	5.00	4.00	2.00
Max	44.00	44.00	43.00	42.00
Range	41.00	39.00	39.00	40.00
Skewness	-0.15	0.19	0.48	0.73
Reliability	0.85	0.84	0.81	0.76
Anchor/Total Correlation	0.87	0.86	0.83	0.80

Table B2. Summary Statistics for New Form Number Correct Score

Non-repeater	Non-repeater	Repeater Group1	Repeater Group2	Repeater Group3
Mean	27.28	23.67	20.54	17.29
SD	7.84	7.82	7.39	6.58
Median	28.00	23.00	20.00	16.00
Min	2.00	5.00	5.00	2.00
Max	44.00	44.00	44.00	41.00
Range	42.00	39.00	39.00	39.00
Skewness	-0.18	0.17	0.44	0.75
Reliability	0.85	0.84	0.82	0.77
Anchor/Total Correlation	0.87	0.87	0.84	0.81

Table B3. Summary Statistics for New Form Number Correct Score with Problematic Anchor

	Non-repeater	Repeater Group1	Repeater Group2	Repeater Group3
Mean	28.29	24.88	21.23	18.22
SD	7.67	7.79	7.38	6.75
Median	29.00	25.00	21.00	17.00
Min	5.00	3.00	2.00	3.00
Max	44.00	44.00	43.00	43.00
Range	39.00	41.00	41.00	40.00
Skewness	-0.30	0.02	0.35	0.65
Reliability	0.85	0.84	0.81	0.78
Anchor/Total Correlation	0.86	0.87	0.85	0.82

Table B4. Summary Statistics for Anchor Test Number Correct Score

Non-repeater	Non-repeater	Repeater Group1	Repeater Group2	Repeater Group3
Mean	7.30	6.22	5.26	4.37
SD	2.67	2.71	2.52	2.28
Median	7.00	6.00	5.00	4.00
Min	0.00	0.00	0.00	0.00
Max	12.00	12.00	12.00	12.00
Range	12.00	12.00	12.00	12.00
Skewness	-0.21	0.10	0.39	0.55
Reliability	0.66	0.65	0.60	0.52

Table B5. Summary Statistics for Problem Anchor Test Number Correct Score

Non-repeater	Non-repeater	Repeater Group1	Repeater Group2	Repeater Group3
Mean	8.35	7.35	6.23	5.23
SD	2.51	2.67	2.67	2.49
Median	9.00	7.00	6.00	5.00
Min	0.00	0.00	0.00	0.00
Max	12.00	12.00	12.00	12.00
Range	12.00	12.00	12.00	12.00
Skewness	-0.56	-0.21	0.09	0.35
Reliability	0.66	0.66	0.64	0.59

Table B6. WRMSB of Equating with Non-problematic Anchor Test

	Mean	SD
Remove Repeaters	0.04	0.03
Remain Repeaters	0.03	0.02
Proportion of repeaters = 0%	0.03	0.02
Proportion of repeaters = 25%	0.04	0.02
Proportion of repeaters = 35%	0.04	0.03
$\theta_{R1} \sim N$ (-0.5, 1)	0.03	0.02
$\theta_{R2} \sim N$ (-1.0, 1)	0.04	0.02
$\theta_{R3} \sim N$ (-1.5, 1)	0.04	0.03
N = 20	0.05	0.03
N = 50	0.04	0.02
N = 100	0.04	0.02
N = 200	0.03	0.02
N = 300	0.03	0.02
N = 400	0.03	0.02
N = 500	0.03	0.02
Circle -arc	0.03	0.03
Identity	0.06	0.01
Rasch equating	0.03	0.02
Nominal weight mean	0.03	0.02
Overall	0.04	0.02

Table B7. WRMSB of Equating with problematic Anchor Test

	Mean	SD
Remove Repeaters	0.04	0.03
Remove Problematic Anchor	0.06	0.03
Remain Repeaters and Anchor	0.09	0.07
Proportion of repeaters = 0%	0.04	0.03
Proportion of repeaters = 25%	0.07	0.04
Proportion of repeaters = 35%	0.09	0.06
$\theta_{R1} \sim N$ (-0.5, 1)	0.06	0.05
$\theta_{R2} \sim N$ (-1.0, 1)	0.06	0.05
$\theta_{R3} \sim N$ (-1.5, 1)	0.07	0.05
N = 20	0.07	0.04
N = 50	0.05	0.04
N = 100	0.06	0.05
N = 200	0.06	0.06
N = 300	0.06	0.05
N = 400	0.06	0.06
N = 500	0.06	0.05
Circle -arc	0.05	0.04
Identity	0.06	0.01
Rasch equating	0.10	0.08
Nominal weight mean	0.06	0.06
Overall	0.06	0.05

Table B8. WSEE: Equating with Non-problematic Anchor Test

	Mean	SD
Remove Repeaters	0.06	0.06
Remain Repeaters	0.06	0.05
Proportion of repeaters = 0%	0.06	0.06
Proportion of repeaters = 25%	0.06	0.06
Proportion of repeaters = 35%	0.06	0.06
$\theta_{R1} \sim N$ (-0.5, 1)	0.06	0.06
$\theta_{R2} \sim N$ (-1.0, 1)	0.06	0.06
$\theta_{R3} \sim N$ (-1.5, 1)	0.06	0.06
N = 20	0.12	0.09
N = 50	0.07	0.06
N = 100	0.08	0.07
N = 200	0.05	0.04
N = 300	0.05	0.04
N = 400	0.04	0.03
N = 500	0.04	0.03
Circle -arc	0.06	0.04
Identity	0.00	0.00
Rasch equating	0.11	0.04
Nominal weight mean	0.01	0.06
Overall	0.06	0.06

Table B9. WSEE: Equating with problematic Anchor Test

	Mean	SD
Remove Repeaters	0.06	0.06
Remove Problematic Anchor	0.07	0.07
Remain Repeaters and Anchor	0.06	0.06
Proportion of repeaters = 0%	0.06	0.06
Proportion of repeaters = 25%	0.07	0.07
Proportion of repeaters = 35%	0.07	0.07
$\theta_{R1} \sim N$ (-0.5, 1)	0.06	0.07
$\theta_{R2} \sim N$ (-1.0, 1)	0.07	0.07
$\theta_{R3} \sim N$ (-1.5, 1)	0.06	0.06
N = 20	0.13	0.11
N = 50	0.08	0.07
N = 100	0.08	0.07
N = 200	0.06	0.05
N = 300	0.04	0.04
N = 400	0.04	0.03
N = 500	0.03	0.03
Circle -arc	0.06	0.04
Identity	0.00	0.00
Rasch equating	0.11	0.04
Nominal weight mean	0.10	0.06
Overall	0.06	0.07

Table B10. WRMSE: Equating with Non-problematic Anchor Test

	Mean	SD
Remove Repeaters	0.08	0.04
Remain Repeaters	0.08	0.05
Proportion of repeaters = 0%	0.08	0.05
Proportion of repeaters = 25%	0.08	0.05
Proportion of repeaters = 35%	0.09	0.05
$\theta_{R1} \sim N$ (-0.5, 1)	0.08	0.05
$\theta_{R2} \sim N$ (-1.0, 1)	0.08	0.05
$\theta_{R3} \sim N$ (-1.5, 1)	0.09	0.05
N = 20	0.14	0.08
N = 50	0.10	0.05
N = 100	0.10	0.05
N = 200	0.07	0.03
N = 300	0.07	0.03
N = 400	0.06	0.02
N = 500	0.06	0.02
Circle -arc	0.07	0.04
Identity	0.06	0.01
Rasch equating	0.11	0.04
Nominal weight mean	0.10	0.06
Overall	0.08	0.05

Table B11. WRMSE: Equating with problematic Anchor Test

	Mean	SD
Remove Repeaters	0.09	0.05
Remove Problematic Anchor	0.10	0.07
Remain Repeaters and Anchor	0.12	0.08
Proportion of repeaters = 0%	0.08	0.06
Proportion of repeaters = 25%	0.11	0.06
Proportion of repeaters = 35%	0.12	0.08
$\theta_{R1} \sim N$ (-0.5, 1)	0.09	0.07
$\theta_{R2} \sim N$ (-1.0, 1)	0.11	0.07
$\theta_{R3} \sim N$ (-1.5, 1)	0.11	0.06
N = 20	0.16	0.10
N = 50	0.11	0.06
N = 100	0.11	0.07
N = 200	0.10	0.06
N = 300	0.09	0.06
N = 400	0.08	0.05
N = 500	0.08	0.05
Circle -arc	0.08	0.05
Identity	0.06	0.01
Rasch equating	0.15	0.07
Nominal weight mean	0.13	0.08
Overall	0.10	0.07

Table B12. CDC: Equating with Non-problematic Anchor Test

	Mean	SD
Proportion of repeaters = 25%	-0.06	0.25
Proportion of repeaters = 35%	-0.08	0.27
$\theta_{R1} \sim N$ (-0.5, 1)	0.00	0.09
$\theta_{R2} \sim N$ (-1.0, 1)	-0.06	0.20
$\theta_{R3} \sim N$ (-1.5, 1)	-0.15	0.37
N = 20	-0.03	0.27
N = 50	-0.10	0.28
N = 100	-0.08	0.24
N = 200	-0.06	0.28
N = 300	-0.08	0.27
N = 400	-0.07	0.21
N = 500	-0.08	0.25
Circle -arc	-0.25	0.23
Rasch equating	0.19	0.20
Nominal weight mean	0.12	0.14
Overall	-0.07	0.26

Table B13. CDC: Equating with Problematic Anchor Test

	Mean	SD
Proportion of repeaters = 25%	-0.82	0.34
Proportion of repeaters = 35%	-0.93	0.41
$\theta_{R1} \sim N$ (-0.5, 1)	-0.76	0.35
$\theta_{R2} \sim N$ (-1.0, 1)	-1.14	0.44
$\theta_{R3} \sim N$ (-1.5, 1)	-1.03	0.39
N = 20	-0.79	0.17
N = 50	-0.71	0.32
N = 100	-0.90	0.27
N = 200	-0.88	0.25
N = 300	-1.05	0.40
N = 400	-1.09	0.44
N = 500	-0.94	0.37
Circle -arc	-0.88	0.28
Rasch equating	-1.44	0.32
Nominal weight mean	-0.79	0.33
Overall	-0.93	0.36

Table B14. DA: Equating with Non-problematic Anchor Test

	Mean	SD
Remove Repeaters	0.87	0.03
Remain Repeaters and Anchor	0.88	0.02
Proportion of repeaters = 0%	0.88	0.02
Proportion of repeaters = 25%	0.87	0.03
Proportion of repeaters = 35%	0.88	0.03
$\theta_{R1} \sim N$ (-0.5, 1)	0.88	0.03
$\theta_{R2} \sim N$ (-1.0, 1)	0.87	0.02
$\theta_{R3} \sim N$ (-1.5, 1)	0.88	0.03
N = 20	0.87	0.05
N = 50	0.88	0.03
N = 100	0.88	0.02
N = 200	0.88	0.02
N = 300	0.88	0.02
N = 400	0.88	0.02
N = 500	0.88	0.02
Circle -arc	0.88	0.02
Identity	0.87	0.02
Rasch equating	0.88	0.02
Nominal weight mean	0.88	0.03
Overall	0.88	0.03

Table B15. DA: Equating with problematic Anchor Test

	Mean	SD
Remove Repeaters	0.87	0.03
Remove Problematic Anchor	0.88	0.02
Remain Repeaters and Anchor	0.87	0.02
Proportion of repeaters = 0%	0.88	0.02
Proportion of repeaters = 25%	0.86	0.02
Proportion of repeaters = 35%	0.87	0.03
$\theta_{R1} \sim N$ (-0.5, 1)	0.87	0.02
$\theta_{R2} \sim N$ (-1.0, 1)	0.87	0.02
$\theta_{R3} \sim N$ (-1.5, 1)	0.88	0.03
N = 20	0.86	0.05
N = 50	0.87	0.02
N = 100	0.88	0.02
N = 200	0.87	0.02
N = 300	0.87	0.02
N = 400	0.87	0.01
N = 500	0.87	0.02
Circle -arc	0.88	0.02
Identity	0.86	0.03
Rasch equating	0.87	0.02
Nominal weight mean	0.87	0.03
Overall	0.87	0.03

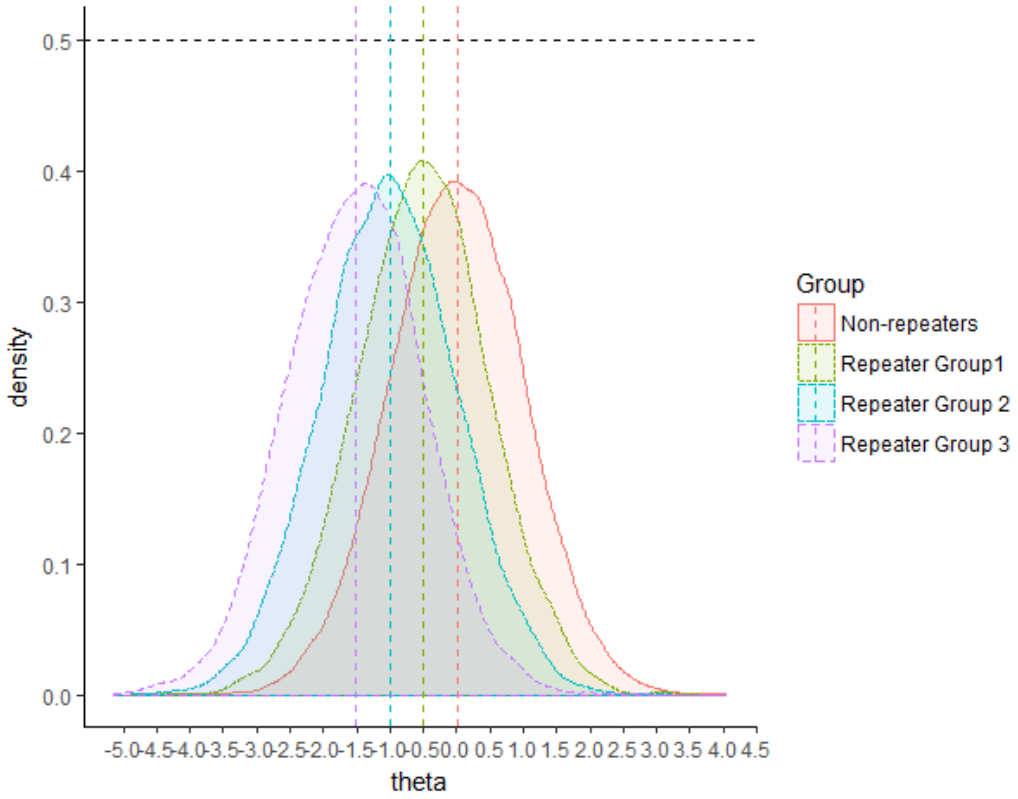


Figure B1. Ability Distribution in the Population

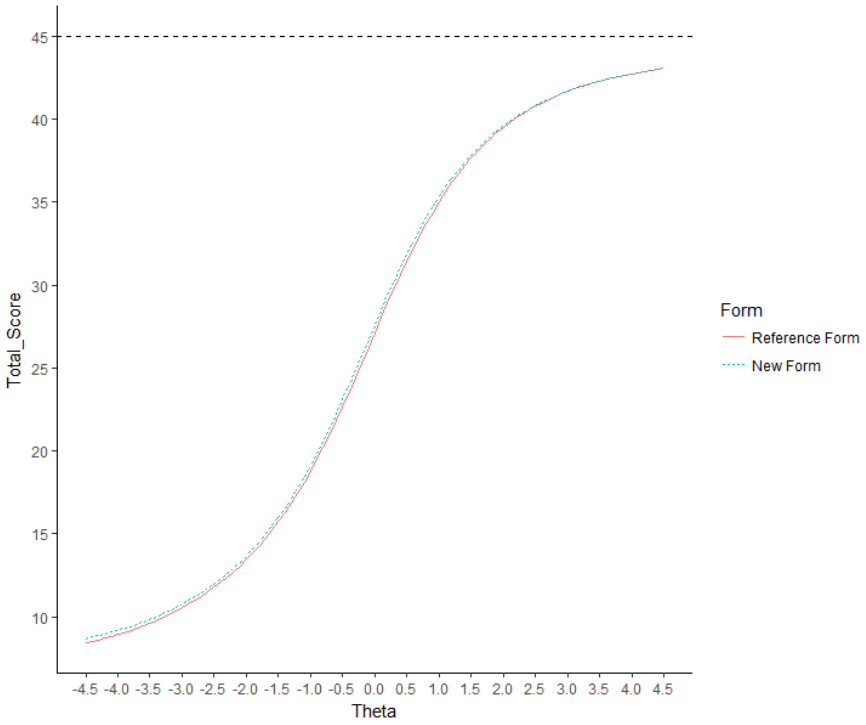


Figure B2. Test Characteristics Curves of Reference and New Form

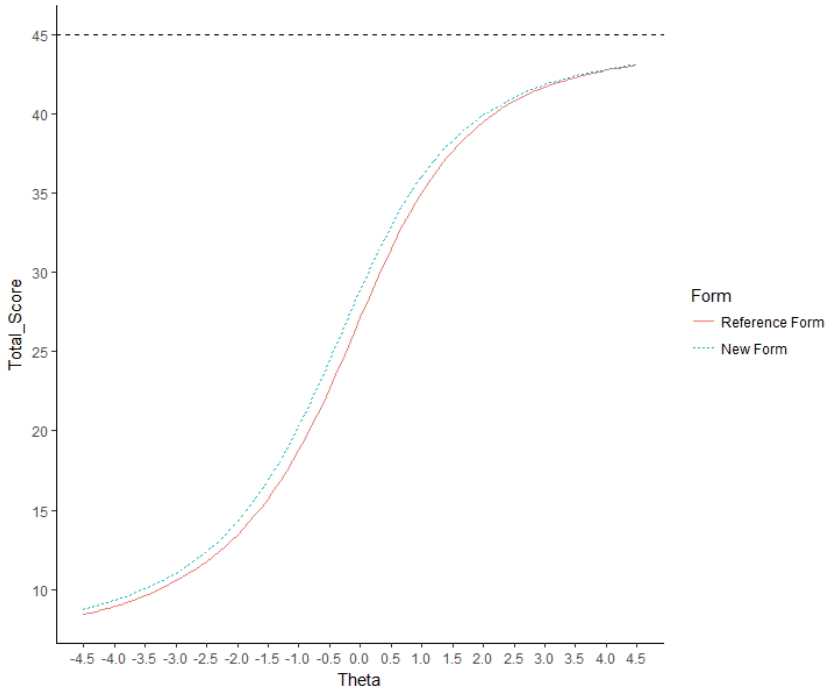


Figure B3. Test Characteristics of Reference and New Form with Problematic Anchor

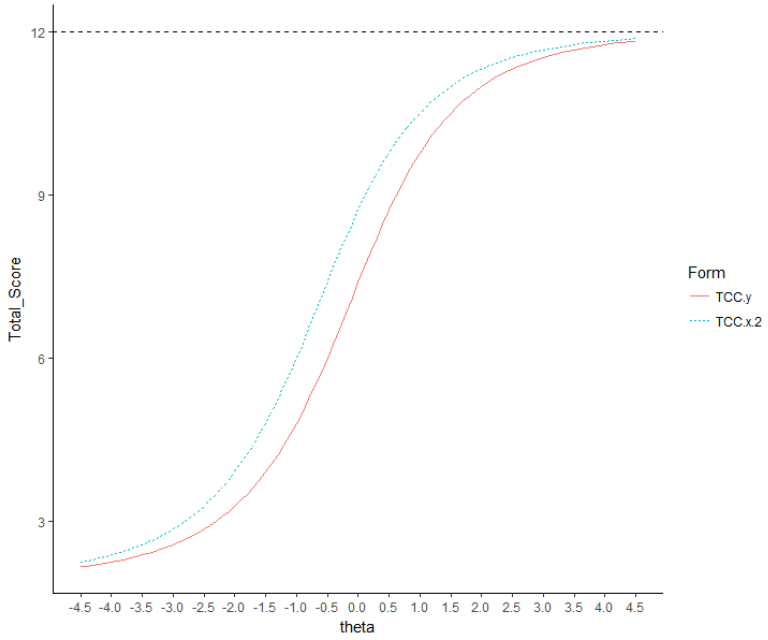


Figure B4. Test Characteristic Curves of Anchor Tests

REFERENCES

Albano, A. D. (2015). A general linear method for equating with small samples. Journal of Educational Measurement, *52*(1), 55-69.

Albano, A. D. (2016). equate: An R package for observed-score linking and equating. *Journal of Statistical Software, 74*(8), 1-36.

Andrulis, R. S., Starr, L. M., & Furst, L. M. (1978). The effect of repeaters on test equating. *Educational and Psychological Measurement*. *38*(2), 341-349.

Angoff, W. H. (1971). *Scales, norms and equivalent scores*. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Babcock, B., Albano. A., & Raymond. M. (2012). Nominal weights mean equating: a method for very small samples. *Educational and Psychological Measurement, 72*(4), 608-628.

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28*(2), 147–162.

Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In F.M. Lord & M. R. Novick, *Statistical theories of mental rest scores* (Part 5, pp. 397-479), Reading, MA: Addison-Wesley.

Bourque, M. L., Goodman, D., Hambleton, R. K., & Han, N. (2004). *Reliability estimates for the ABTE tests in elementary education, professional teaching knowledge, secondary mathematics and English/language arts (Final Report)*. Leesburg, VA: Mid-Atlantic Psychometric Services

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9-49). New York: Academic.

Cai, L. (2013). flexMIRT ®version 2: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. *Chapel Hill, NC: Vector Psychometric Group*.

Chen, W. H., Lenderking, W., Jin, Y., Wyrwich, K. W., Gelhorn, H., & Revicki, D. A. (2014). Is Rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? An example using PROMIS pain behavior item bank data. *Quality of Life Research*, *23*(2), 485-493.

Dorans, N. J. (2004). Equating, concordance, and expectation. *Applied psychological measurement*, *28*(4), 227- 246.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equitability of tests: Basic theory and the linear case. Journal of Educational Measurement, 37, 281-306.

Dorans, N. J., Liu, J., & Hammond, S. (2008). Anchor test type and population invariance: An exploration across subpopulations and test administrations. *Applied Psychological Measurement*, *32*(1), 81-97

Duong, M. Q., & von Davier, A, A. (2012). Observed-score equating with a heterogeneous target population. *International Journal of Testing*, *12*(3), 224-251. doi: 10.1080/15305058.2011.620725

Gianopulos, G. (2008). *The robustness of Rasch true score preequating to violations of model assumptions under equivalent and nonequivalent populations* (Doctoral dissertation, University of South Florida). Retrieved from http://scholarcommons.usf.edu/cgi/viewcontent.cgi?article=1258&context=etd

Gorham, J. L., & Bontempo, B. D. (1996). *Repeater patterns on NCLEX using CAT versus NCLEX using paper-and-pencil testing*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Haebara, T. (1980). Equating logistic ability scales by a weighted least square method. *Japanese Psychological Research, 22*(3), 144-149.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Hanson, B. A., & Beguin, A. A. (2002). Obtaining a common scale for item response theory parameters using separate and concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26*(1), 3–24.

Hanson, B. A., Zeng, L., & Colton, D. A. (1994). A comparison of presmoothing and postsmoothing methods in equipercentile equating. *American College Testing Program, 1994*(4).

Harris, D. J., & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education, 6*, 195-240.

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied psychological measurement, 20*(2), 101-125.

Holland, P. W., & Thayer, D. T. (1987). Notes on the use of log-linear models for fitting discrete probability distributions. *ETS Research Report Series, 1987*(2).

Kim, H., & Kolen, M. J. (2010). The effect of repeaters on equating. *Applied Measurement in Education*, *23*(3), 242-265.

Kim, S., & von Davier. A. A., & Haberman. S., (2011). Practical application of a synthetic linking function on small-sample equating. *Applied Measurement in Education*, *24*(2), 95-114. doi: 10.1080/08957347.2011.554601

Kim, S., & Walker, M. E. (2012). Investigating Repeaters Effects on Chained Equipercentile Equating With Common Anchor Items. *Applied Measurement in Education*, *25*(1), 41-57. doi: 10.1080/08957347.2012.635481

Kim, S., Livingston. S. A., & Lewis. C. (2011). Collateral information for equating in small samples: a preliminary investigation. *Applied Measurement in Education*, *24*(4), 302-323. doi: 10.1080/08957347.2011.607057

Kim. S., & Livingston., S. A. (2010). Comparisons among small sample equating methods in a common-item design. *Journal of Educational Measurement*, *47*(3), 286-298.

Kim. S., von Davier. A. A., & Haberman. S. (2008). Small-sample using a synthetic linking function. *Journal of Educational Measurement*, *45*(4), 325-342.

Kolen, M. J. (2004). Population invariance in equating and linking: Concept and history. *Journal of Educational Measurement*, *41*(1), 3-14.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York: Springer.

Liao, C., & Qu, Y. (2010). *Alternate forms test-retest reliability and test score changes for the TOEIC speaking and writing tests* (ETS internal report). Princeton, NJ: Educational Testing Service.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch measurement transactions, 7*(4), 328.

Livingston, S. A., & Kim, S. (2009). The circle-arc method for equating in small samples. *Journal of Educational Measurement, 46*(3), 330-343.

Livingston, S. A., & Kim. S. (2008). Small-sample equating by the circle-arc method. *ETS Research Report Series, 2008*(2). Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2008.tb02125.x/pdf

Livingston. S. A. (1993). Small-sample equating with log-linear smoothing. *Journal of Educational Measurement, 30*(1), 23-39.

Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Routledge.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement, 8*(4), 453-461. doi:10.1177/014662168400800409

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using Rasch model. *Journal of Educational Measurement, 17*(3), 179-193.

Macro, G. L. (1977). Item characteristic solutions to three intractable testing problems. *Journal of Educational Measurement, 14*(2), 139-160.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174. doi:10.1007/bf02296272

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series*, *1992*(1).

Ogasawara, H. (2001). Least squares estimation of item response theory linking coefficients. *Applied Psychological Measurement*, *25*(4), 373-383.

Parshall, C. G., Houghton, P. D. B., & Kromrey, J. D. (1995). Equating error and statistical bias in small sample linear equating. *Journal of Educational Measurement*, *32*(1), 37-54.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). *Scaling, norming, and equating*. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 221-262).

Puhan, G. (2009). What effect does the inclusion or exclusion of repeaters have on test equating? *ETS Research Report Series*, *2009*(1). Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2009.tb02176.x/pdf

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rizopoulos, D. (2006). ltm: An R Package for Latent Variable Modelling and Item Response Theory Analyses, *Journal of Statistical Software*, *17*(5), 1-25.

Rogers, W, T., & Radwan, N. (2015). Impact of inclusion of varying percentages of repeaters on equating. *International Journal of Testing*, *15*(3), 177-192. doi: 10.1080/15305058.2014.976865

Samejima, F. (1972). *A general model for free-response data* (Psychometric Monograph No. 18). Richmond, VA: Psychometric Society. Retrieved from https://www.psychometricsociety.org/sites/default/files/pdf/MN18.pdf

Skaggs, G. (2005). Accuracy of random groups equating with very small samples. *Journal of Educational Measurement*, *42*(4), 309-330.

Stage, C., & Ögren, G. (2004). *The Swedish scholastic assessment test (SweSAT). Development, results and experiences* (EM No.49). Umeå, Sweden: Umeå University, Department of Educational Measurement.

Stocking, M. L., & Lord, F. M. (1983) Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*(2), 201-210.

Stone, M., & Yumoto, F. (2004). The effect of sample size for estimating Rasch/IRT parameters with dichotomous items. *Journal of applied measurement*, *5*(1). 48

Sunnassee, D. (2011). *Conditions affecting the accuracy of classical equating methods for small samples under the NEAT design: A simulation study* (Unpublished doctoral dissertation). The University of North Carolina at Greensboro.

Thornton, A. E., Stilwell, L. A., & Reese, L. M. (2006). *The validity of law school admission test scores for repeaters: 2001 through 2004 entering law school classes* (LSAC Research Report No. 06-02). Newtown, PA: Law School Admission Council.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). The Chain and post-stratification methods for observed-score equating and their relationship to population invariance. *Journal of Educational Measurement*, *41*(1), 15-32.

Wang, T., Lee, W. C., Brennan, R. L., & Kolen, M. J. (2008). A comparison of the frequency estimation and chained equipercentile methods under the common-item nonequivalent groups design. *Applied Psychological Measurement, 32*(8), 632-651.

Weeks, J. P. (2010). plink: An R Package for Linking Mixed-Format Tests Using IRT-Based Methods. *Journal of Statistical Software, 35*(12), 1-33.

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. New York: Springer.

Yang, W. L., Bontya, A. M., & Moses, T. P. (2011). Repeater effects on score equating for a graduate admission exam. *ETS Research Report Series, 2011*(1). Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2011.tb02253.x/pdf